1999

Quantile estimation using auxiliary information with applications to soil texture data

Pamela Joy Abbitt
Iowa State University

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Quantile estimation using auxiliary information with applications to soil texture data

by

Pamela Joy Abbitt

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Major Professors: F. Jay Breidt and Sarah M. Nusser

Iowa State University

Ames, Iowa

1999

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Pamela Joy Abbitt

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Overview

The National Cooperative Soil Survey (NCSS) is a collaborative program involving the USDA and a state agency, often the state's Agricultural Experiment Station. The NCSS program is charged with constructing reports for each county which contain soil maps and descriptions of characteristics for all soils found within the county. These maps are periodically updated through the NCSS program to provide current information on characteristics for different soils. Updates are based on soil surveys involving extensive field work. This information is used by contractors, farmers and others for land use planning purposes and by scientists to develop models based on soil characteristics.

Descriptions of map units in a soil survey typically include variables such as the color, texture and structure of the soil. A key concept in these descriptions is the representative value. A representative value for a particular soil characteristic is often calculated as the midpoint of an observed range of values or the mode of the observed values. Both of these descriptive measures can be heavily biased by the use of purposive sampling designs. The objectives of the pilot project include finding statistically sound methods of estimating distributional quantities for many soil characteristics.

A variety of sampling methods have been applied in soil surveys, where many soil characteristics are described for a large geographic area, and soil science research, where a few soil characteristics are investigated for a relatively small geographic area. Although soil science research has a stronger tradition of using probability sampling designs, data on the distribution of soil properties are typically gathered during soil surveys using purposive sampling methods.

Probability sampling designs have been considered impractical for soil surveys. However, recent developments in GIS (Geographic Information System) and GPS (Global Positioning System) technologies have made it feasible for such designs to be implemented. A GIS can be used to facilitate sample selection for complex designs. Also, GPS technology allows soil scientists to quickly locate randomly selected sites for data collection. Traditional and alternative sampling designs are described in Chapter 2.

In a pilot project in MLRA 107 in western Iowa, a stratified multi-phase sampling plan was used to conduct soil survey updates in two counties. In general, multi-phase designs are used when a variable of interest is expensive to measure, but it is strongly related to another variable which is inexpensive to observe. The inexpensive variable is called the auxiliary variable. In the pilot project, laboratory measurements are the expensive variables of interest and related field measurements are relatively inexpensive. The objective of a multi-phase design is to measure the auxiliary variable on a relatively large sample and the variable of interest on a small sample. In the estimation stage, the auxiliary information is used to improve estimators of distributional quantities relating to the variable of interest. In particular, we consider estimation of quantiles incorporating auxiliary information.

Sampling units for the pilot project are points, which we refer to as *sites*. A Markov chain sampling design which encourages geographic spread was used to create a list of sites in each county. The list of sites is stratified using a GIS and digitized soil maps from a previous survey. A stratified systematic sample was selected from the Markov chain sample. Samples for subsequent phases were systematic subsamples of the first phase sample. The sampling design is described in Chapter 2 and in Abbitt and Nusser (1995). Some examples of preliminary analyses of data collected under the multi-phase design are also given in Chapter 2 and in Abbitt, Breidt and Nusser (1997) and Abbitt, Goyeneche and Schumi (1998).

Estimation of quantiles of the distribution of soil texture for varying depths in a profile is an important and challenging problem. Soil texture is an important consideration in land use and management. In the pilot project, field and laboratory determinations of

texture are made for each horizon at selected sites. A *horizon* is roughly a layer of soil. The soil texture data structure is quite complicated due to the horizon-specific measurements and multi-phase sampling design. For example, full profile data are available at some sites, while only surface horizon data are available at other sites. In Chapter 3, we describe the structure of the available soil texture data.

Two approaches to estimating quantile profiles are considered for the soil texture data. First, a semi-parametric imputation-based approach was developed to exploit the multi-phase structure of the sampling design. In the analysis of data collected under a multi-phase sampling design, estimators which incorporate auxiliary information are often used. This is an area of considerable research in survey sampling. For example, ratio estimators and regression estimators are two well-known examples. See, for example, Särndal et al. (1991).

In particular, Chambers and Dunstan (1986) presented an estimator for a distribution function which incorporates auxiliary information. Auxiliary information is assumed to be available for each element in a finite population, while the variable of interest is available only for a sample. A fitted model for sampled elements and the values of the auxiliary variable for non-sampled elements are used to improve upon the naive estimator, which uses only sampled values of the variable of interest.

Chambers and Dunstan (1986) conjectured a partial linearization of a quantile estimator derived from their distribution function estimator. However, it is possible to derive a complete linearization of this estimator. The linearized expression can be used to derive an expression for the asymptotic variance of the quantile estimator. In Chapter 4, we present results for the quantile estimator as well as a small simulation study which compares the asymptotic variance to the empirical variance.

The Chambers and Dunstan quantile estimator motivates a semi-parametric imputation-based approach to estimating quantile profiles of laboratory determinations of soil texture for the pilot project (Chapter 5). Auxiliary information is available in the form of field determinations of soil texture. Due to the multi-phase sampling design used for data collection, field determinations are available at more sites than laboratory deter-

minations and field determinations for surface horizons are available at more sites than full profiles of field data.

A second approach to estimating quantile profiles of soil texture is to use a hierarchical model as described in Chapter 6. A hierarchical model is used to describe the relationships among the observed data and unknown parameters. The model incorporates information about the horizons observed at each site. This information is not used explicitly in the imputation approach of Chapter 5. Prior distributions for all parameters in the model can be specified in such a way that draws from the full posterior distribution can be simulated via a standard Markov Chain Monte Carlo (MCMC) technique. These draws are used to produce estimated quantile profiles which are comparable in interpretation to those from the imputation approach. In addition to estimated quantile profiles, this procedure provides estimated distributions for the quantile profiles, for all parameters in the model and for any transformation of these parameters.

The imputation approach employs models which appear to be reasonable to impute missing data. Simulation results suggest that jackknife methodology may be a valid way to estimate the variance of the quantile estimator. However, under explicit distributional assumptions, the hierarchical model provides richer output. In addition, the Bayesian approach seems to provide a more comprehensive framework, parts of which can be used for analyzing other variables collected in the pilot project.

Data can be simulated from the hierarchical model presented in Chapter 6. We analyze the simulated data using the imputation approach to evaluate its performance. A large source of bias in the imputation approach is discovered and an improvement is proposed. Chapter 7 presents the results of this simulation and a general comparison of the two approaches.

1.2 Dissertation Organization

Chapter 2 reviews sampling methods which have been used by soil scientists and presents the sampling design which was implemented in a pilot project in MLRA 107 in western Iowa. This chapter was prepared to be submitted to the *Soil Science Society of*

America Journal. Chapter 3 describes the structure of the soil texture data collected in the MLRA 107 project and defines notation for use in later chapters.

Chapter 4 presents some results for the quantile estimator derived from the Chambers and Dunstan (1986) distribution function estimator. This quantile estimator is modified for use in analyzing the soil texture data in Chapter 5. Chapter 6 presents a hierarchical model which can also be used for analyzing the soil texture data. The two analysis approaches are compared in Chapter 7. Chaptere 8 contains a brief conclusion.

2 STATISTICAL SAMPLING APPROACHES FOR SOIL SURVEY UPDATES

A paper to be submitted to the Soil Science Society of America Journal

P. J. Abbitt, S. M. Nusser, T. E. Fenton, P. Cowsert and J. Hempel

2.1 ABSTRACT

Information from soil surveys is increasingly needed to support analyses in the areas of soil quality and environmental research and management. To provide a foundation for statistical estimation, statistical sampling procedures should be used to locate sites for collecting soils information. In this paper, we describe statistical tools for selecting samples for soil survey updates. We review sampling methods currently being used and propose a multi-phase sampling plan. A pilot project in which such a design was implemented is described. Example analyses from this type of design are presented.

2.2 INTRODUCTION

The National Cooperative Soil Survey (NCSS) program has prepared detailed soil maps for more than 600 million hectares (ha) in the U.S. Soil surveys are used in many fields that rely on knowledge of the location and characteristics of soil map units (SMUs). These surveys have traditionally been used in farming, forestry, rangeland management, conservation planning, and urban activities such as transportation planning, and residential and industrial development. Updates are usually conducted every 20 to 30 years to adjust boundaries of delineations, to modify descriptions of existing SMUs, and to

define new SMUs.

Information from soil surveys is increasingly needed to support analyses in the areas of soil quality and environmental research and management. Natural resource scientists need point-specific data and statistical descriptions of soil characteristics in order to model complex processes that are affected by these characteristics. Traditional soil survey users can also benefit from more complete map unit descriptions (Brown and Huddleston, 1991).

To provide a foundation for statistical estimation and point-level modeling, statistical sampling procedures should be used to locate sites for collecting soils information. Procedures for randomly selecting points for data collection avoid biases created when points are selected by the data gatherer. Standard sampling protocols such as stratification of the survey area (e.g., by SMU) can be used to create representative samples for the region being surveyed. In addition, sampling and estimation procedures can be applied that balance the need for information at many points against the costs associated with recording many variables at sample points.

In this paper, we describe statistical tools for selecting samples for soil survey updates. We review sampling methods used in the U.S. and other countries to conduct soil surveys and research investigations, and propose a multi-phase sampling plan for use in conducting soil survey updates. A pilot project is described that is designed to test this approach in Crawford and Woodbury Counties in Iowa as part of the Major Land Resource Area (MLRA) 107 soil survey update.

2.3 SAMPLE DESIGNS FOR SOIL SCIENCE

2.3.1 Sample Design Concepts for Soil Survey Updates

Sampling strategies are often compared in terms of how well they estimate a parameter, such as the mean of a particular soil characteristic. One sample design is said to be more efficient than another if the variance of the estimator for the first design is smaller than that of the second. In other words, a better sampling strategy is expected to yield

more precise estimates for the same sample size.

The phrase "random sampling" is often used in soil science as a synonym for simple random sampling (e.g., Wilding, 1985). We use this phrase to include a much larger class of sampling designs [see e.g., Cochran (1977), Thompson (1992), Schreuder et al (1993)] and to distinguish it from purposive sampling. In a purposive sampling design, sampling locations are chosen by the data gatherer using knowledge of the survey area. Random sampling, in the statistical sense, is based on a predefined set of rules to obtain randomly selected locations of sampling units (e.g., points or transects). Random sampling designs include stratified samples, multi-phase samples, and multi-stage cluster samples. In the procedure described below, we use statistical tools such as stratification and multi-phase sampling to gain statistical efficiency while choosing a total sample size that is consistent with the operational resources of more traditional purposive procedures.

Stratification is commonly used to improve the precision of estimates and to ensure that the sample is adequately dispersed across subpopulations, such as SMUs, within a survey area. It is frequently a feature of designs used in soil surveys and soil science research. The population (e.g., the survey area) is divided into mutually exclusive classes, or strata, that partition the population into groups within the survey area (e.g., SMU). Samples are chosen independently in each stratum. More precise estimators can be obtained when the variation of the characteristic of interest within each stratum is small relative to the variation between strata. In addition, if strata define subpopulations of interest, an allocation scheme can be applied to ensure that a sufficient number of sample units are located within each stratum for making inferences on these subpopulations.

Like any type of survey, soil surveys are subject to operational constraints involving schedules, budgets, and personnel. The total cost of a soil survey is determined not only by the number of sites to be visited, but also by the measurements to be taken at each site. In soil surveys, measurements are often taken on properties of the landscape at the site, and visual and chemical determinations may be made on each of several different horizons. More locations can be sampled if the measurements have a low cost, for example, if only visual inspection of the surface horizon is required. However, in

many cases, more expensive measurements, such as chemical determinations for each horizon, are key variables of interest.

It is often reasonable to assume that there are relationships among measurements made at a given site. For example, one expects some characteristics of the surface horizon to be associated with those of deeper horizons for most soils. A multi-phase sample design can be used to minimize data collection efforts by taking advantage of relationships between inexpensive and expensive variables. In multi-phase sampling, inexpensive information is collected at a set of sampling units for the first phase. For the second phase, a smaller sample of units is selected at which both the inexpensive variables and the more expensive variables of interest are observed (Thompson, 1992). Third and subsequent phases may be added. When there are only two phases, two-phase sampling is also called double sampling.

For illustration, consider a two-phase sample design. In phase 1 of a sample, inexpensive measurements such as surface horizon observations can be made on a large set of points. In phase 2, a smaller set of points is selected. For these points, we record the phase 1 measurements and the more costly data items such as a full profile description and lab determinations. Phase 2 points are typically a subset of the phase 1 points. Using the phase 2 points, we estimate relationships between the inexpensive variables and the costly variables. These relationships are then combined with all of the phase 1 data and used to make inferences about the costly variables for the entire study area.

Multi-stage sampling can be useful if there is a hierarchical structure in the population of interest, e.g., points within an area or delineations within a map unit. For this reason, multi-stage designs are often called hierarchical designs in soil science. They involve more than one stage of sampling in which a cluster of units is initially selected, then a subset of units within the sample cluster is selected, and so on (Thompson, 1992). For example, in a two-stage design, the primary sampling unit (PSU) is selected at the first stage. A PSU might be defined as a delineation of a SMU or a selected region such as a 10 m $\times$ 10 m area. The secondary sampling unit (SSU) is selected from within the PSU. For example, a point or points may be selected within the area designated as

the PSU. Data may be collected from the second stage unit only or from both sampling units (e.g., land use for the PSU and a profile description at the point).

2.3.2 Selecting Point Samples for Soil Surveys

For soil surveys, the population of interest is generally the land area within a particular region, such as an MLRA or a county. Soil survey data are often collected at the point level, e.g., characteristics of the soil at a particular point. The area of interest contains an infinite population of points. Because of this, we cannot create a list of all points in the region from which to draw a sample as can often be done with a population of individuals. However, we can still select a sample from the infinite population of points by selecting a finite set of coordinates within a region.

In order to obtain a sample which is representative of the whole area, sample points should be geographically dispersed. A common procedure for selecting a well-dispersed sample is systematic sampling. In a two-dimensional systematic sample, a single point is randomly selected that represents the corner of a grid that covers the study area. Sample points are located at the intersections of the selected grid. Grid samples are considered purposive when the grid is aligned with existing map grid lines such as sections lines instead of randomly selecting the first point. Non-systematic designs (e.g., multi-stage sampling) may also be used in conjunction with stratification to ensure a geographically spread sample.

2.3.3 Purposive Sampling Methods in Soil Surveys

Guidelines for surveying soils in the U.S. are given in the Soil Survey Manual (USDA SCS, 1993). Soil surveying in the U.S. usually relies on purposive point or transect sampling. In the process of mapping soils, observations are made at locations where visible features suggest that the dominant soil or important inclusions will be best expressed (USDA SCS, 1993). Though some guidelines may be provided, the specific location of these sites is at the discretion of the soil scientist.

In transecting, soils are observed at intervals along planned lines of travel. It is

simpler if intervals are regular, but periodicities in the landscape should be avoided (Webster and Cuanalo, 1975). The direction of transects is often chosen by the soil scientist to encounter the most variation within the delineation (Schellentrager, 1991). However, where the pattern of soils is not easily predictable, transect orientation may be chosen at random from a restricted set of directions (USDA SCS, 1993).

The direction of a transect may be selected randomly, but if the orientation is such that no significant variation in the landscape is encountered, the advantage of transecting is defeated. In one method of "random transecting" designed to address this problem, a collection of potential transects is identified. The direction of these transects are chosen to encounter the greatest amount of variation, as in purposive transecting. Observations are then made on a randomly selected sample from the designated transects (Schellentrager, 1991). Because the collection of potential transects is chosen purposively, this is still a purposive sampling design.

Purposive point and transect methods have also been used in other countries such as New Zealand (McLaren and Cameron, 1990), England (Landon, 1984), the Netherlands (Steur et al., 1961), the Philippines (U.N., 1972), and Russia (Nikol'skii, 1963; Revut and Rode, 1969). In many countries, it is recommended that soil scientists choose representative sites for observation. Landon (1984) refers to this as "judgment sampling." McLaren and Cameron (1990) call this "free traversing" since the locations of observations are not restricted by a pre-determined pattern.

A more methodical approach is often recommended for selection of sampling locations if the visual features of the landscape do not help the surveyor plot map unit boundaries. A collection of transects may be used to collect data. Observations may be made along the diagonal of a field or along contours, while avoiding the edges of the area (Revut and Rode, 1969). Nikol'skii (1963) and McLaren and Cameron (1990) suggest other modified transect approaches.

The methods described above are all purposive and rely on the soil scientist's knowledge of the area to select locations. Webster and Oliver (1990) warn that "selection of the 'typical' is not a safe way to sample." In the field, sample sites may be chosen which

support the soil scientist's assessment of the soils and landforms in a delineation. The full range of values in an SMU may not be observed when sites are chosen in this way. Analysis of survey data collected from purposive procedures are frequently implemented as if points were selected randomly (e.g., Beckett and Webster, 1971). However, if a purposive sampling design is used, it is difficult to verify the assumptions required to construct statistically valid estimates of the characteristics of SMUs. Because the full range of values may not be observed, selecting typical or representative sites can produce biased descriptions of SMUs, particularly when the parameter of interest is a range or percentile.

2.3.4 Random Sampling Methods in Soil Surveys

The need for random sampling procedures in soil survey has been acknowledged (Webster, 1977). Random sampling methods have been applied using both point and transect methods in soil survey, although these methods have largely been rejected because of the amount of work involved. Many authors claim that random point procedures require larger sample sizes (e.g., Landon, 1984), although this is not necessarily true. Another concern is that more work is needed to locate randomly selected observation sites (de Gruijter, 1985). In the past, soil scientists have estimated distances from visible features on aerial photographs and then used a compass to traverse to sites. However, random sampling procedures have recently become more practical with the advent of Global Positioning System (GPS) technology.

For more detailed surveys, systematic sampling is sometimes recommended (McLaren and Cameron, 1990). These samples may be aligned with section lines or the national grid in England (e.g., Avery, 1990; Agbu and Olson, 1990). However, systematic samples have not been recommended for general soil surveys, because sample sizes are thought to be prohibitive (Doolittle et al., 1988). From a statistical perspective, systematic sampling across an area is generally more efficient than simple random sampling if observations close together are expected to be similar. However, periodicities in the landscape can result in systematic bias.

Geospatial techniques have been used to select optimal sample designs. For example, Domburg et al. (1994) and Domburg et al. (1997) use estimated variograms to predict sampling error. The predicted sampling error can be used to compare the efficiency of different sampling designs. Estimates of variograms can also be used to determine optimal spacing for a two-dimensional systematic sample (Di et al., 1989; Burgess et al., 1981; Campbell, 1978; McBratney and Webster, 1983; Webster, 1985). However, these methods require previous point information for each SMU or a preliminary study to estimate variograms in order to determine optimal designs.

An alternative approach is to use restricted randomization. Brus et al (1992) studied different methods for updating soil surveys in the Netherlands. They suggest using a two-step procedure in which the first step is a stratified random sample of points within strata defined by map units, where points are allocated proportional to the surface area of the map unit. This sample is used to estimate variances of phosphate sorption characteristics within each map unit. The variance estimates are then used to augment the stratified sample so that the final allocation across map units is optimal with respect to the variance within map units.

2.3.5 Random Sampling Methods in Soil Science Research

While random sampling methods have had limited use in soil surveys, they have been widely implemented in soil science research. Many research projects have narrow objectives involving a restricted area, a small number of SMUs, or only a few soil characteristics. In contrast to soil surveys, these investigations are often conducted with the specific purpose of obtaining statistical descriptions of selected soil characteristics.

Studies in the area of soil science research often employ two-dimensional systematic sampling schemes similar to those discussed in the previous section (Di et al., 1989; Khan and Nortcliff, 1982), multi-stage designs (Nortcliff, 1978; Aljibury and Evans, 1961; Edmonds et al., 1985; Hammond et al., 1985) or a mixture of both designs (Youden and Mehlich, 1937; de Gruijter, 1985). Multi-phase designs have also been used. Gessler et al. (1995) used this type of approach to relate expensive soil characteristics to more

easily observed environmental variables that were available for the entire study area.

2.4 THE MLRA 107 PILOT PROJECT

2.4.1 Overview of Sample Design

As part of the MLRA 107 soil survey update, a sampling plan based on the statistical tools described above was developed for Crawford and Woodbury Counties in Iowa. The design approach was identical in both counties, although sample sizes were not. For simplicity, we present the application of this design in Crawford County, Iowa.

The sample design is a stratified multi-phase design. To obtain a set of point coordinates, a dense sample was drawn across the entire area using a Markov chain sampling procedure which will encourage geographic spread without using a grid (Breidt, 1995a). Abbitt and Nusser (1995) and Breidt (1995b) outline the application of this procedure to soil survey updates. The area is first divided into a grid of contiguous non-overlapping rectangles, each of which is quite small. One point is selected in each of these rectangles, resulting in a geographically dispersed sample. The location of the points are chosen using a method which discourages selection on a regular grid. The resulting sample is an equal probability sample.

We use this approach to develop what we call the phase 0 sample, from which sample points are selected for site visits. For each site in the phase 0 sample, the map unit symbol (MUS) of the delineation in which it is located is obtained using a geographic information system (GIS) and digitized soil maps from the previous soil survey. The number of phase 0 sites lying in a particular SMU is approximately proportional to the acreage of the SMU. The MUS recorded for phase 0 sample sites is used to define strata in selecting subsamples for subsequent phases. Allocations of the sample sizes across SMUs can be controlled in a manner that provides adequate sample sizes for SMUs or groups of SMUs.

A phase 1 sample is selected from the phase 0 sample for each SMU. Landscape information and surface horizon data are collected at each phase 1 site. The phase

2 sample is a subsample consisting of 25% of the phase 1 sites, at which full profile descriptions are recorded. Half of the phase 2 sites are selected for a phase 3 sample; at each phase 3 site, soil is collected from each horizon for laboratory determinations. In areas where significant shifts in soil characteristics or changes in classification have occurred, supplemental sites can also be selected to increase the information available for boundary determinations.

Soil scientists locate points in the sample using GPS units in combination with mylar overlays depicting the sampled coordinates for a small area and aerial photographs to provide context for the mylar overlays.

In the following sections, we provide a detailed description of the procedures used to select the Crawford County sample.

2.4.2 Sample Selection

A very large sample of points was selected to develop a basis for selecting samples within each SMU. This large sample is the phase 0 sample. To select the phase 0 sample, a rectangle that completely surrounded Crawford County was divided into a 245 $\times$ 245 grid of rectangles, each about 200 $\times$ 170 m. Using the Markov chain sampling procedure described in Breidt (1995b), a sample of 60,025 sites was selected. Sites lying outside the county were discarded, leaving 53,557 phase 0 sites. ARC/INFO and ArcView were used to link each point to the MUS of its delineation as recorded in digitized soil maps for Crawford County created from the previous survey. Boundaries for SMUs in the digital coverage were judged to be a good approximation to current SMU boundaries.

A stratified multi-phase sample was selected from the phase 0 sample. Strata were defined to be SMUs. The total sample size for Crawford County was determined to be approximately the same number of sites that would have been selected using traditional soil survey update procedures. Sample sizes were allocated to each stratum (SMU) depending on the area in the county covered by the SMU in digital soil map. The

allocation rule was defined to be

$$n_k = \begin{cases} .027A_k^{.70} & \text{if } A_k > 2,024 \text{ ha for SMU } k \\ 4 & \text{otherwise,} \end{cases} \tag{2.1}$$

where n_k is the number of phase 1 sites for SMU k and A_k is the area of SMU k in hectares. This rule ensures that small SMUs are allocated a minimum of four sites, and that large SMUs have proportionately more sites. Fig. 2.1 shows phase 1 sample sizes in relation to SMU size.

Phase 1 sites were selected as a stratified systematic subsample of phase 0 sites. First, sites were ordered by latitude and longitude within each SMU. Then, for each SMU, a random start was determined, and points were selected at regular intervals from the list of points for the SMU. The number of points selected in each SMU was determined by the allocation rule in Eq.[2.1].

Even though systematic sampling was used, the spatial distribution of the sample is not systematic. This is because locations in the phase 0 sample were randomly selected according to the Markov Chain procedure and because delineations for the same SMU are not contiguous. An important feature of this design is that it is an equal probability design within a stratum. This means that within SMUs, all points have an equal probability of being selected.

The sample sizes for phase 2 and 3 were determined based on the ability to complete field work within resource and timing constraints for the project. To obtain the phase 2 sample for each SMU, a systematic subsample of one-fourth of the phase 1 for the SMU points was selected for phase 2. One-half of the phase 2 sites for each SMU were systematically selected for phase 3. Only one phase 2 site was selected for the SMUs containing only four phase 1 sites. These SMUs are paired with the SMU that has the next largest area, and one phase 3 site is chosen from the two phase 2 sites belonging to the pair. Table 2.1 shows a small portion of the full list of phase 0 sample sites for Crawford County and an example sample selected using this procedure. Fig. 2.2 shows the dispersion of actual sample sites in Crawford County for each phase. Fig. 2.3 shows an example of selected locations for the three phases in a two square mile (two section)

area of Crawford County.

2.4.3 Supplemental Points

The three-phase design assumes that the delineations from the previous soil survey are approximately correct. The proposed design can also be augmented when changes in soil conditions require additional information to define new delineation boundaries. This can happen, for example, when soil classification schemes are revised or when a natural phenomenon that alters soil composition has occurred since the previous soil survey. In MLRA 107, both of these circumstances exist in the bottomlands. In addition to the reclassification of some soil series, 1993 floods are expected to have had a significant effect on the location of soils near river beds in Iowa. To address these situations, a sample of supplemental points, referred to as soil symbol points, was drawn in the affected region. Only a map unit symbol (MUS) is recorded for the soil symbol points.

The supplemental sample for Crawford County contained 561 soil symbol points selected systematically from the phase 0 sample using the following rules. In bottomland SMUs comprising less than 405 ha,

$$n_{sk} = \begin{cases} 14 + .0028(A_k) & \text{if } A_k < 405 \text{ ha} \\ 84 & \text{otherwise,} \end{cases}$$

where n_{sk} is the number of soil symbol points that were selected in addition to the phase 1 points selected for SMU k, and A_k is the area of SMU k in hectares. These rules were selected to generate sample sizes that are roughly equivalent to those obtained if transect methods had been used. Fig. 2.2 shows the distribution of soil symbol sites across Crawford County, and Fig. 2.3 shows an example of selected locations for the supplemental sample in two sections of Crawford County.

2.4.4 Data Analysis

Because data collection is still underway for the pilot project, final analyses can not be reported. To provide a better understanding of the types of analyses that can be generated from a statistical approach to soil survey updates, we present examples to

demonstrate how multi-phase information can be used in estimating soil parameters. Preliminary analyses are also presented which provide a feel for more complex analyses.

Suppose we wish to estimate μ_Y, the average clay content of the B horizon for a particular SMU. For our purposes, we ignore any error in the observation due to the measurement process and assume that the true clay content of a horizon can be observed. In order to observe this variable, a soil scientist would need to probe to the depth of the B horizon. Time or resource constraints may only allow collection of this data at a small sample of sites. This small sample is the phase 2 sample. For the soil, we are investigating, the clay content of the A horizon is expected to be highly correlated with clay content of the B horizon. Since these measurements are less expensive to collect, the soil scientist can afford to collect this data at all of the phase 2 sites and another sample of sites. The phase 1 sample is the collection of all sites where the clay content of the A horizon is recorded. The phase 2 sample is a subset of the phase 1 sample.

Let Y_i be the clay content in the B horizon and X_i be the clay content in the A horizon for site i. Suppose n sites were selected with equal probability from the SMU in phase 1, and X_i is measured for each site in the sample. For phase 2, m sites were selected with equal probability from phase 1 sites, and Y_i is recorded in addition to X_i.

A simple estimator of μ_Y is the sample mean of the B horizon clay content from the m phase 2 sites. This estimator can be written as

$$<Y_2>= m^{-1} \sum_{i=1}^{m} Y_i. \tag{2.2}$$

Note that only m observations are available to calculate this estimate. If there is a strong relationship between X and Y, we can use this relationship and the larger set of X measurements to improve upon the estimator in Eq.[2.2].

Assume that clay content in the B horizon is a multiple of clay content in the A horizon; a plausible model is

$$Y_i = rX_i. \tag{2.3}$$

In two-phase sampling, we estimate r from the phase 2 sample of m sites where both X and Y are available. An estimator of r is the ratio of the phase 2 sample mean of Y and

the phase 2 sample mean of X, given by

$$\hat{r} = <Y_2> / <X_2>, \tag{2.4}$$

where $<X_2> = m^{-1} \sum_{i=1}^{m} X_i$. Based on model Eq.[2.3] and the estimator for r in Eq.[2.4], an alternative estimator of the average clay content in the B horizon is

$$\hat{\mu}_Y = \hat{r} <X_1>, \tag{2.5}$$

where $<X_1> = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean of the clay content in the A horizon from the n points in the phase 1 sample. Note that both $<X_1>$ and $<X_2>$ are estimates of the mean of X, but $<X_1>$ is a more precise estimator since it is based on more observations.

In estimating μ_Y, we do not have the luxury of n observations of Y. The estimators $<Y_2>$ and $\hat{\mu}_Y$ both estimate μ_Y, the average clay content in the B horizon. However, $\hat{\mu}_Y$ in Eq.[2.5] exploits the relationship between X and Y and the "extra" information contained in the larger sample of X values by using $<X_1>$, the more precise estimator of the mean of X. It can be shown that $\hat{\mu}_Y$ will be more precise than $<Y_2>$ if there is a strong correlation between X and Y [Cochran (1977), Thompson (1992)].

Figure 2.4 demonstrates how the estimator $\hat{\mu}_Y$ is constructed graphically. The horizontal axis represents values of X, the proportion of clay in the A horizon. The vertical axis represents values of Y, the proportion of clay in the B horizon. The dark solid line indicates the estimated relationship between X and Y, $Y = \hat{r}X$. The slope of this line is $\hat{r}$, which is estimated from the phase 2 sample means. The simple estimate of μ_Y, $<Y_2>$, is adjusted via the estimated model. Graphically, this means that the point $(<X_1>, \hat{\mu}_Y)$ must fall on the line $Y = \hat{r}X$. If, as in Fig. 2.4, $<X_1>$ is smaller than $<X_2>$ and $\hat{r}$ is positive, then $\hat{\mu}_Y$ is smaller than $<Y_2>$.

In some instances, it may be more meaningful to estimate the proportion of area that has clay content in the B horizon above a certain percentage, p. We can use the procedure described above to estimate this parameter by defining new variables for site

i in the SMU as follows:

$$X_i' = \begin{cases} 1 & \text{if } X_i > p \\ 0 & \text{if } X_i \leq p \end{cases}$$

and

$$Y_i' = \begin{cases} 1 & \text{if } Y_i > p \\ 0 & \text{if } Y_i \leq p \end{cases}$$

where Y_i' indicates whether the clay content in horizon B is above p and X_i' indicates whether the clay content in horizon A is above p for site i. The phase 2 sample means, $<X_2'>$ and $<Y_2'>$, now represent the proportion of the phase 2 sample with clay content greater than p for the A and B horizons, respectively. Formulas analogous to those in Eqs.[2.4] and [2.5] are used to calculate $\hat{r}', <X_1'>$, and $\hat{\mu}_Y'$. In this case, $\hat{\mu}_Y'$ will be the estimated proportion of the SMU acreage with B horizon clay content exceeding p percent.

It should be noted that in the MLRA 107 pilot project, as might be expected in any update, SMU boundaries from the old survey are not the same as in the new survey. Thus, the resulting samples within the new SMUs are no longer equal probability samples. In this situation, sampling weights are required. For example, in estimating the average clay content in horizon B weighted means are used to estimate $<X_1>$, $<X_2>$ and $<Y_2>$, e.g., $<X_1>= \sum_{i=1}^{n} w_i X_i$ where w_i is the sampling weight.

Sampling weights are commonly used in survey sampling to account for unequal selection probabilities among sites. A sampling weight is a measure of the area represented by a sample point. For example, if 10 sites are randomly selected from a 200 ha region, the sampling weight is 200/10 = 20 ha. However, if 40 sites are selected from the 200 ha region, the weight would be smaller (5 ha). Known information about an area can be incorporated into sampling weights. In the MLRA 107 pilot project, we intend to incorporate the new area for each SMU in the weights using a ratio approach similar to that just described [Cochran (1977); Thompson (1992)]. This approach ensures that estimates calculated from the sample data reflect published area totals for SMUs.

These examples are designed to provide a simple illustration of how phase 1 information can be used to obtain more precise estimates of quantities related to variables that can only be measured at phase 2 sites. In research currently underway, regression estimators are being developed to produce summary statistics for inclusion in the soil survey report. Parameters being estimated include means, variances, and percentiles as well as standard errors of the estimates for characteristics within each SMU. These kinds of statistical summaries provide information on the central tendency and variability of properties for a SMU as well as a measure of the precision of the estimate.

Two examples of analyses currently being investigated are presented below. The details of the analyses are omitted here. However, details of some preliminary analyses have been presented in Abbitt et al. (1997) and Abbitt et al. (1998). The results are provided to indicate the type of estimates that can be produced for a statistical sample.

Table 2.2 provides estimates of slope gradient percentiles for different slope classes and erosion phases of a particular soil series. The slope classes present are B(2 to 5%), C (5 to 9%), D (9 to 14%) and E (14 to 20%). The slope class letter is accompanied by a number indicating the erosion phase. If no number is present, the erosion phase is 'none to slight'. The numbers 2 and 3 represent 'moderate' and 'severe' erosion, respectively.

Each column of Table 2.2 corresponds to one phase of the series (identified by a unique slope class and erosion phase combination). Each row represents an estimated percentile. For example, in this series, 50% of the land area classified as a class C map unit with moderate erosion is estimated to have a slope gradient less than or equal to 6.8%. This is the 50th percentile, often called the *median*. Alternatively, 10% of the area classified as belonging to phase B of the map unit is estimated to have a slope gradient less than 2.7%. Tables similar to Table 2.2 provide detailed information about the variation of a soil property in relation to the phase of a SMU.

Figure 2.5 shows estimates of the 10th and 90th percentile of clay content in relation to depth for a particular map unit. The horizontal axis represents values of clay content. The vertical axis represents depth from the surface in cm. The top of each plot corresponds to the surface, while the bottom corresponds to approximately 120 cm

(48") below the surface. The solid lines represent the estimated percentile profiles for clay. The dashed lines represent each profile plus and minus two standard errors. The interval between the dashed lines is an approximate 80% confidence interval for the estimated percentile profile. A wider interval indicates that the estimate is less precise. For example, the estimate of the 10th percentile is less precise at 120 cm below the surface than at 50 cm below the surface.

Profiles representing other estimated percentiles can be produced. A collection of estimated percentile profiles describes how the distribution of a texture component changes over depth. Distributions can be compared across map units using these estimates. Percentile estimates can also be summarized by master horizon designation instead of by inches.

2.5 SUMMARY

In order to obtain statistically valid estimates, we must randomly sample from the entire population of interest. Numerous statistical sampling designs have been created for this purpose that address both scientific objectives and operational constraints. Recent technologies such as GPS and GIS have made it possible to develop a method of selecting statistical point samples for soil survey updates that are based on realistic sample sizes and that take advantage of the existing knowledge of the location of soils and their characteristics.

The specific objective of this sample design is to provide information on the distribution of soil characteristics for all map units in a survey area, while balancing the intensity of field work with the need to collect adequate information to support statistical estimation. It can be applied in other parts of the country with different soils and classifications and with alternative objectives. Although the selection procedure relies heavily on GIS, analog procedures can be applied to accomplish the same goals.

We are continuing to develop statistical methods for summarizing soil composition and properties of SMUs using soil survey data collected from multi-phase designs. A major contribution of this approach is that data collected under this design can be used

for a broader suite of statistical analyses than in a more traditional soil survey update. The data can be used to obtain alternative estimates of representative values and ranges that are statistically defensible. In addition, alternative estimates of other measures such as means, percentiles, or parametric distributions can be generated along with estimated standard errors. The database resulting from this approach also supports geographically-linked modeling efforts.

2.6 ACKNOWLEDGEMENTS

This work has been supported in part by cooperative agreement 68-3A75-5-72 between the USDA Natural Resources Conservation Service and Iowa State University. The authors wish to thank Jay Breidt, Bennie Clark, Louis Boeckman, Wayne Fuller, Gary Medlyn, and Tom Reedy for their extensive contributions to this research. The project was partially funded under the leadership of Dennis Lytle and Richard Arnold. Craig Ditzler and Ellis Benham were involved in the initial design. Richard Lensch, Sam Steckley, and others have been involved in field data collection.

2.7 REFERENCES

Abbitt, P. J. and S. M. Nusser. 1995. Sampling approaches for soil survey updates. Proc. of the Am. Stat. Assoc., Sect. on Stat. and the Environment, p. 87-91.

Abbitt, P. J., J. J. Goyeneche and J. Schumi. 1998. An approach to estimating clay profile distributions. Proc. of the Am. Stat. Assoc., Sec. on Survey Research Methodology. p. 372-377.

Abbitt, P. J., S. M. Nusser and F. J. Breidt. 1997. A nonlinear two-phase predictor for soil survey updates. Proc. of the Am. Stat. Assoc., Sec. Surv. Res. Methodology. p. 657-660.

Agbu, P. A. and K. R. Olson. 1990. Spatial variability of soil properties in selected Illinois Mollisols. Soil Sci. 150(5):777-786.

Aljibury, F. K. and D. D. Evans. 1961. Soil sampling for moisture retention and bulk density measurements. Soil Sci. Soc. of Am. Proc. 25:180-183.

Avery, B. W. 1990. Soils of the British Isles. University Press, Cambridge.

Beckett and R. Webster. 1971. Soil variability: A review. Soils and Fertilizer 34:1-15.

Breidt, F. J. 1995a. Markov chain designs for one-per-stratum sampling. Survey Methodology 21(1):63-70.

Breidt, F. J. 1995b. Markov chain designs for one-per-stratum spatial sampling. Proc. of the Am. Stat. Assoc., Sect. on Surv. Res. Methods 1:356-361.

Brown, R. B. and J. H. Huddleston. 1991. Presentation of statistical data on map units to the user. In Mausbach, M. J. and L. P. Wilding, (eds.) Spatial Variabilities of Soils and Landforms, SSSA Special Publication #28.. SSA Inc., Madison, WI.

Brus, D. J., J. J. deGruijter and A. Breeuwsma, 1992. Strategies for updating soil survey information: A case study to estimate phosphate sorption characteristics. J. of Soil Sci. 43:567-81.

Burgess, R. Webster and A. B. McBratney, 1981. Optimal interpolation and isarithmic mapping of soil properties IV: Sampling strategy. J. of Soil Sci. 32:643-659.

Campbell, J. B. 1978. Spatial variation of sand content and pH within single contiguous delineations of two soil mapping units. Soil Sci. Soc. Am. J. 42:460-464.

Cochran, G. W. 1977. Sampling Techniques. New York: Wiley and Sons.

de Gruijter, J. J. 1985. Transect sampling for reliable information on mapping units. In Soil Spatial Variability, ISSS and SSSA, PUDOC.

Di, H. J., B. B. Trangmar and R. A. Kemp. 1989. Use of geostatistics in designing sampling strategies for soil survey. Soil Sci. Soc. Am. J. 53:1163-1167.

Domburg, P., J. J. deGruijter and D. J. Brus. 1994. A structured approach to designing soil survey schemes with prediction of sampling error from variograms. Geoderma 62:151-164.

Domburg, P., J. J. deGruijter and P. vanBeck. 1997. Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. Geoderma 75:183-201.

Doolittle, J. A., R. A. Rebertus, G. B Jordan, E. I. Swenson and W. H. Taylor. 1988. Improving soil-landscape models by systematic sampling with ground-penetrating radar. Soil Surv. Horizons 29(2).

Edmonds, W., Baker, J., and Simpson, T. W. 1985. Variance and scale influences on classifying and interpreting soil mapping units. Soil Sci. Soc. Am. J. 49:957-961.

Food and Agriculture Organization of the U.N. 1972. Soil Fertility Survey and Research: The Philippines, vol. : Soil survey and land classification of the Penaranda river irrigation system area. United Nations Development Program, Rome.

Gessler, P. E., I. D. Moore, N. J McKenzie, and P. J. Ryan. 1995. Soil-landscape modeling and spatial prediction of soil attributes. Int. J. of Geographic Information Systems 9(4):421-432.

Hammond, L. C., W. C. Pritchett and V.Chew. 1985. Soil sampling and soil heterogeneity. Soil Sci. Soc. of Am. Proc. 22:548-552.

Khan, M. A. and S. Nortcliff. 1982. Variability of selected soil micronutrients in a single soil series in Berkshire, England. J. of Soil Sci. 33:763-770.

Landon, J. R., editor 1984. Booker Tropical Soil Manual. Booker-Tate Publishing, London.

McBratney, A. B. and R. Webster, 1983. Optimal interpolation and isarithmic mapping of soil properties V: Co-regionalization and multiple sampling strategies. J. of Soil Sci. 34:137-162.

McLaren, R. G. and K. C. Cameron, 1990. Soil Science: An introduction to the properties and management of New Zealand soils. Oxford Univ. Press, Auckland.

Nikol'skii, N. N. 1963. Practical Soil Science. Israel Program for Scientific Translation, Jerusalem.

Nortcliff, S. 1978. Soil variability and reconnaissance soil mapping: a statistical study in Norfolk. J. of Soil Sci. 29:403-418.

Revut, I. B. and A. A. Rode. editors. 1969. Experimental Methods of Studying Soil Structure. Kolos Publishers, Leningrad.

Schellentrager, G. W. 1991. Map unit transects and statistical analysis of transect data. Proc. Iowa Cooperative Soil Survey Scientists Workshop.

Schreuder, H. T., T. G. Gregoire and G. B. Wood. 1993. Sampling Methods for Multi-resource Forest Inventory. Wiley, New York.

Soil Survey Division Staff. 1993. Soil Survey Manual. USDA SCS, handbook no. 18. http://www.statlab.iastate.edu/soils/nssh/.

Steur, G. G. L. et al. 1961. Methods of soil surveying in use at the Netherlands soil survey institute. Boor en Spade 11:59-67.

Thompson, S. 1992. Sampling. Wiley, New York.

Webster, R. 1977. Quantitative and numerical methods in soil classification and survey. Clarendon Press, Oxford.

Webster, R. 1985. Quantitative Spatial Analysis of Soil in the Field, volume 3, chapter 1. Springer-Verlag.

Webster, R. and H. E. Cuanalo de la C. 1975. Soil transect correlograms of north Oxfordshire and their interpretations. J. of Soil Sci. 26(2):176.

Webster, R. and M. A. Oliver, 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press.

Wilding, L. P. 1985. Spatial variability: its documentation, accommodation and implication to soil surveys. In Soil Spatial Variability, Wageningen. ISS and SSSA, PUDOC.

Youden and Mehlich. 1937. Selection of efficient methods for soil sampling. Boyce Thompson Inst. for Plant Res. 9:59-70.

Table 2.1 Section of the phase 0 sample list from a SMU with 16 phase 0 sites. To select a total of eight points, a phase 1 site is randomly selected, then sites are selected at fixed intervals in the list from this point. A similar process is conducted for subsequent phases.

SMU	Longi-tude	Lati-tude		Selected in Phase 1		Selected in Phase 2		Selected in Phase 3
AdD3	−95.3270	41.8680						
AdD3	−95.3643	41.8794	→	×				
AdD3	−95.4010	41.8864						
AdD3	−95.4087	41.8901	→	×				
AdD3	−95.4186	41.9062						
AdD3	−95.4207	41.9187	→	×	→	×	→	×
AdD3	−95.4224	41.9196						
AdD3	−95.6051	41.9226	→	×				
AdD3	−95.3028	41.9378						
AdD3	−95.3268	41.9412	→	×				
AdD3	−95.2349	41.9566						
AdD3	−95.3634	41.9581	→	×				
AdD3	−95.3357	41.9652						
AdD3	−95.4560	41.9712	→	×	→	×		
AdD3	−95.2306	41.9989						
AdD3	−95.4920	42.0050	→	×				

Table 2.2 Estimated percentiles of slope gradient for several phases of a series.

Percentile	*B*	*C*	*C*2	*D*	*D*2	*D*3	*E*2	*E*3
10th	2.7	4.4	4.0	6.8	7.0	6.7	10.0	10.1
30th	3.2	5.5	5.8	8.6	9.1	9.5	12.4	12.8
50th	3.5	6.2	6.8	9.6	10.3	10.0	13.8	14.4
70th	3.7	6.8	7.8	10.6	11.5	12.3	15.1	15.9
90th	4.1	7.6	9.0	11.8	12.9	14.1	16.8	17.9

Fig. 2.1. Sample allocation across SMUs. Plot depicts sample size within a SMU in relation to the acreage of the SMU.

Fig. 2.2. Map of point locations in Crawford County for Phases 1, 2 and 3 and for the supplemental point sample. Underlying grid represents section lines.

Fig. 2.3. Two sections in Crawford county. Each polygon is a delineation. Black areas make up the soil map unit MoB. Gray areas are bottomlands in which extra samples were selected. Sample points are denoted according to the legend below.

Fig. 2.4. Graphical representation of the estimator . The naive estimator is adjusted according to the estimated model. The estimator $\hat{\mu}_Y$ is more precise than $<Y_2>$ if X and Y are strongly correlated.

Fig. 2.5. Estimates of the 10th and 90th percentile profiles for a series. The dashed lines represent approximate 95% confidence intervals for each estimate.

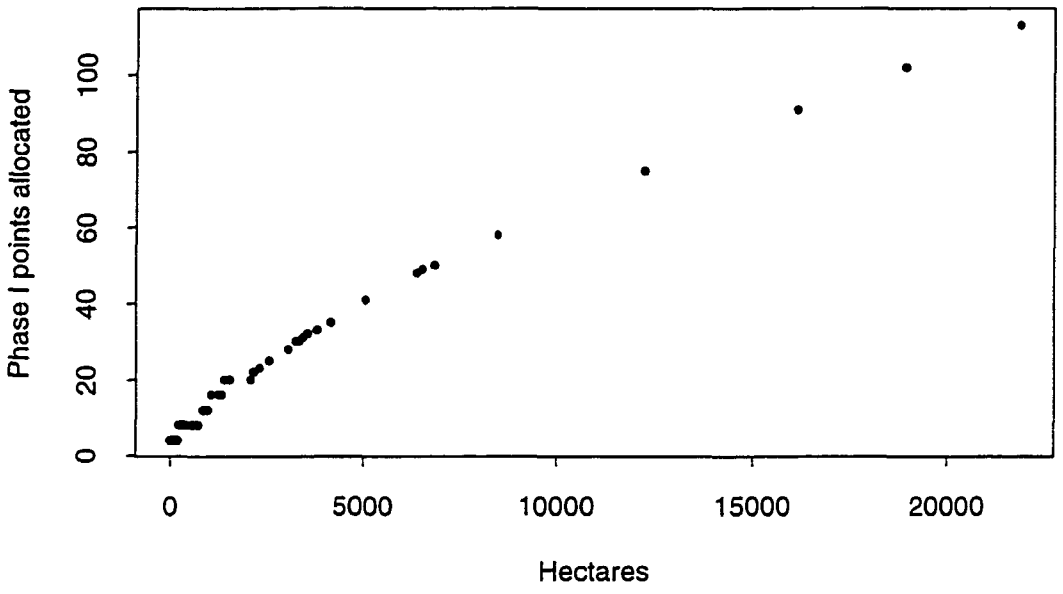
Phase I points allocated
0
20
40
60
80
100
0
5000
10000
15000
20000
Hectares

Figure 2.1

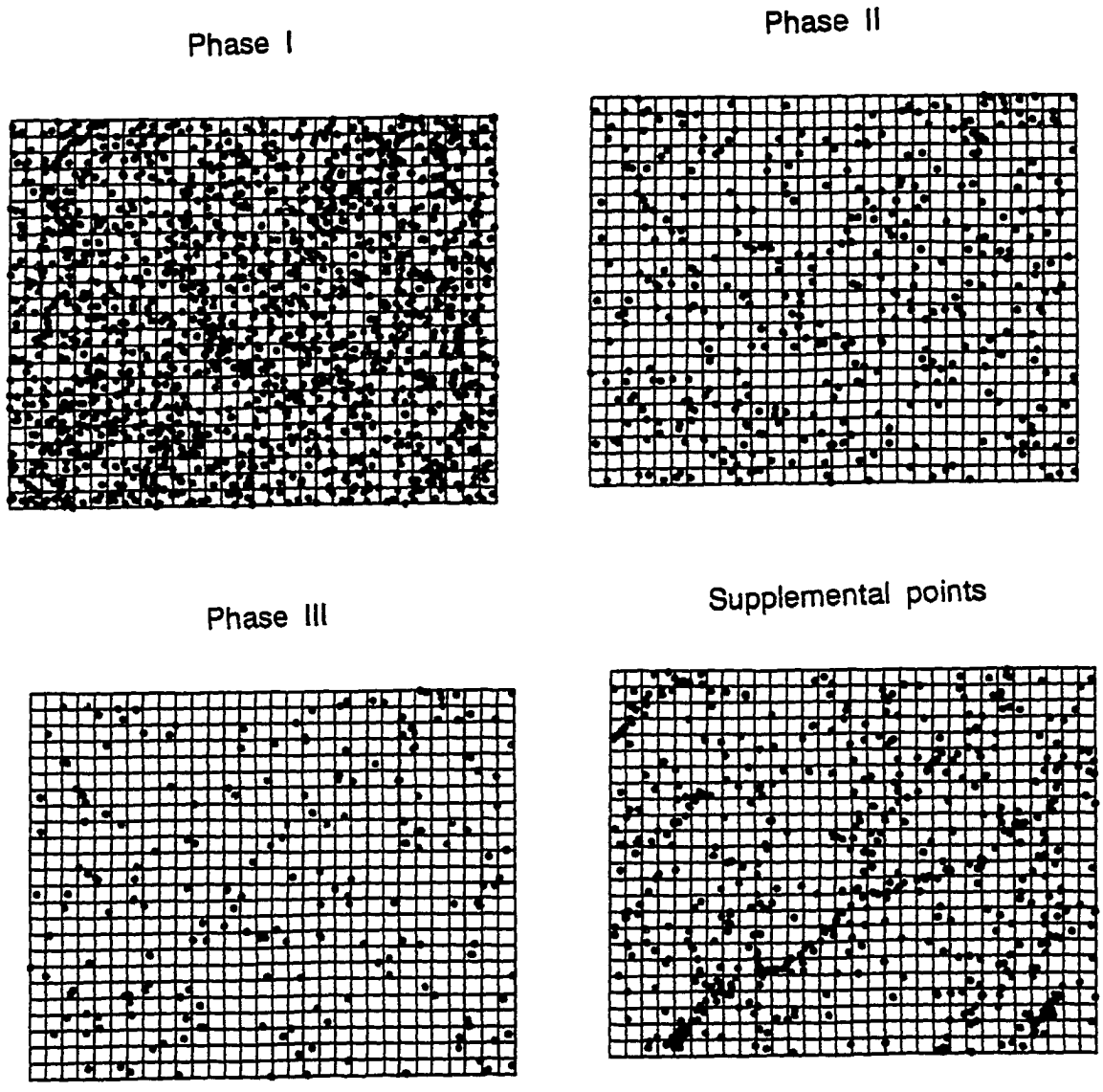

Figure 2.2

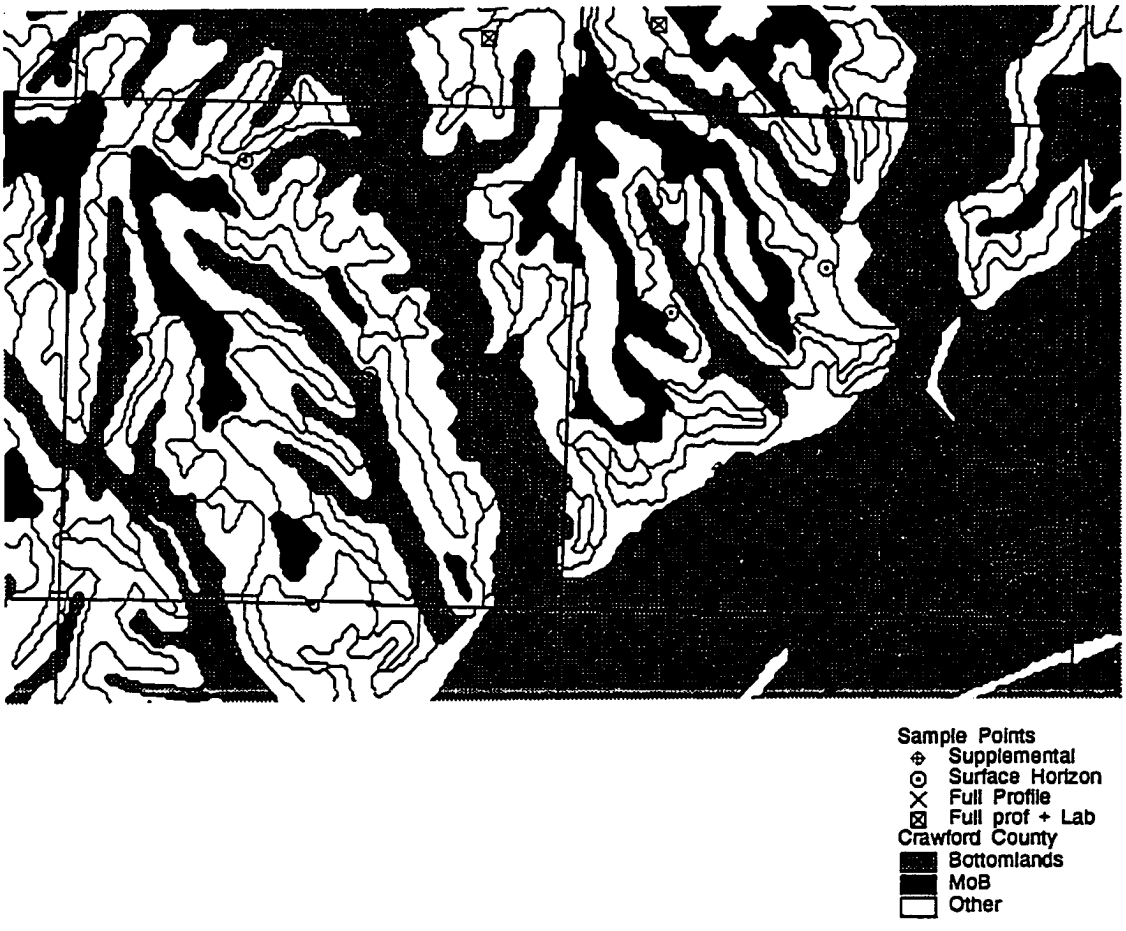
Sample Points
Supplemental
Surface Horizon
Full Profile
Full prof + Lab
Crawford County
Bottomlands
MoB
Other

Figure 2.3

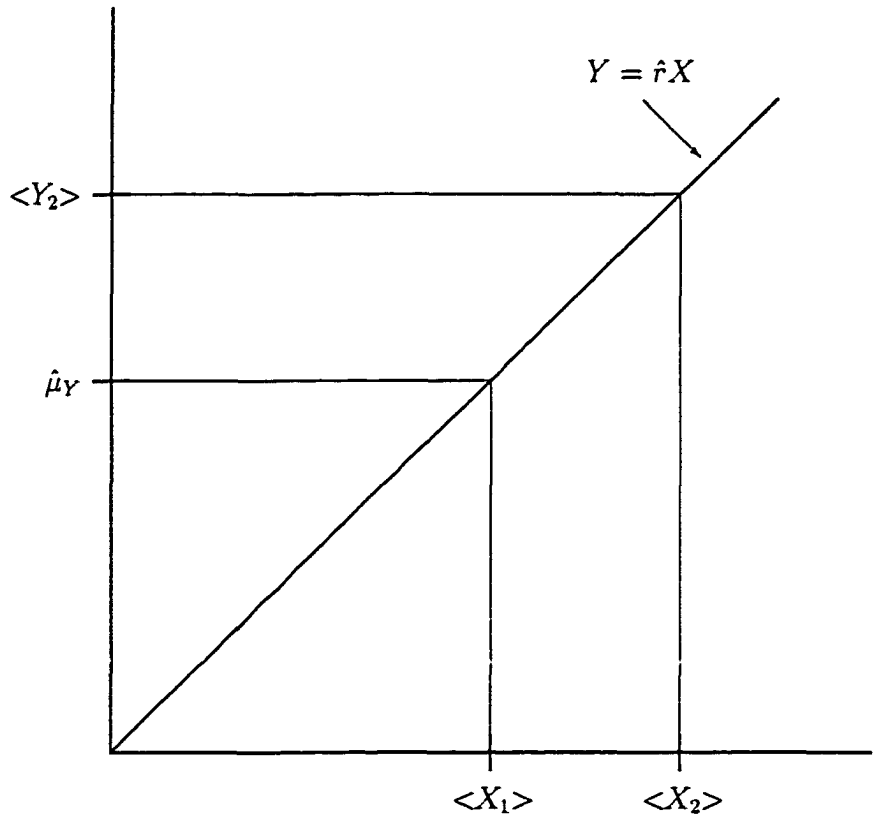

Figure 2.4

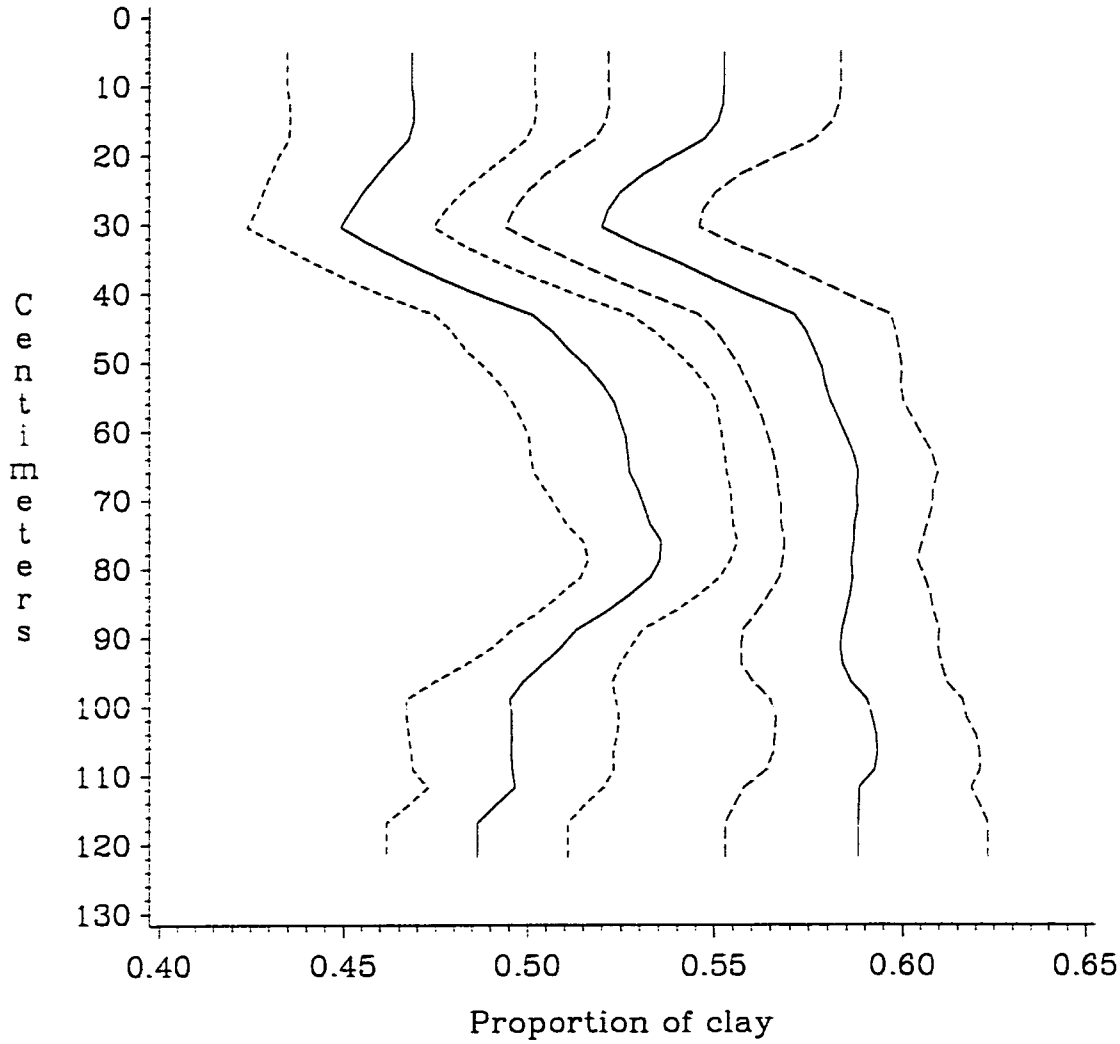
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Centimeters
0.40
0.45
0.50
0.55
0.60
0.65
Proportion of clay

Figure 2.5

3 SOIL TEXTURE DATA AND ANALYSIS OBJECTIVES

3.1 Introduction

Soil texture is an important consideration in land use and management. At a particular site, the texture of the soil may change with depth. We wish to investigate the distribution of soil texture profiles for a collection of map units. In particular, we wish to obtain estimates of quantile profiles and corresponding standard errors.

This chapter contains a description of soil texture data collected for the MLRA 107 pilot project described in Section 2.4 and used as the basis for the analyses in Chapters 5 and 6. Terminology and notation are presented throughout this chapter for use in both analyses. A review of important features of the sampling design used in the pilot project is given in Section 3.2. Available data are described in Section 3.3. Section 3.4 describes relevant variables observed in the pilot project. In both analysis approaches, a transformation of the texture data is used. This transformation is presented in Section 3.5.

3.2 Sampling design and data collection overview

The study population in the MLRA 107 pilot project is a county. Map units or collections of map units are domains of interest for soil survey updates. A map unit refers to all areas within a county which have been identified as a particular soil with the same slope class and erosion phase. Sampling units are points on the land which are dimensionless. Thus the population is an infinite collection of sampling units. We call these sampling units *sites*.

The sampling design used in the pilot project consists of three phases. In all phases,

variables are recorded by horizon. A *horizon* is a layer of soil which differs from the adjacent layers in physical, biological or chemical properties. The thickness, labels and order of horizons are likely to differ from site to site.

For the first phase sample sites, information is collected on the physical characteristics that are easily determined from the *surface horizons.* The surface horizons are the top one or two horizons at the site. For second phase sample sites, field-observable data are collected on all horizons to a depth of 48 inches, where possible. This type of description of soil characteristics as they vary across depth is called a *profile.* In the third phase sample, laboratory determinations are made on soil samples taken from each horizon. For more details of the sampling design, see Section 2.4.

3.3 Data structure

Because data collection is still underway and monior deviations in protocol occurred, the design phases (phases 1, 2 and 3) do not accurately reflect the data which are currently available. We will use $\mathcal{S}$ to refer to sites where field data are only available for the surface horizon. The set of sites with profiles of field data to a depth of 48 inches will be denoted $\mathcal{F}$. We will use $\mathcal{L}$ to refer to sites where profiles to a depth of 48 inches of field and laboratory data are available.

The sets $\mathcal{S}$, $\mathcal{F}$ and $\mathcal{L}$ are defined as mutually exclusive sets. That is, no sites are contained in more than one set. Let $\mathcal{D}$ denote $\mathcal{S} \cup \mathcal{F} \cup \mathcal{L}$. The set $\mathcal{D}$ is the collection of all sites in the phase 1 sample for which some amount of data has been collected. Note that no laboratory data are available for sites in $\mathcal{S} \cup \mathcal{F}$. The set $\mathcal{L}$ is not the same as the phase 3 sample because laboratory data on all phase 3 sites has not yet been received. Also, laboratory data were collected at some additional sites. We use the notation $|\mathbb{S}|$ to denote the size of a set $\mathbb{S}$.

Laboratory data have been collected for 133 of the 244 phase 3 sites for which the field data are available. Laboratory measurements have been collected for 11 additional sites which were not orginally selected in the phase 3 sample. The number of sites and horizons for which data are currently available is summarized in Table 3.1. A total

of 3567 sites were selected for the phase 1 sample. However, due to some sites being unsuitable for data collection (e.g., the site falls on a road, in a ditch, etc.), we expect a total of approximately 3000 phase 1 sample sites. Thus, approximately two-thirds of the final data are being used in the current analysis.

Table 3.1 Number of sites and horizons in the current data set.

Set	Number of Sites *	Number of Horizons **
$\mathcal{S}$	1520	1783
$\mathcal{F}$	334	1973
$\mathcal{L}$	144	882
$\mathcal{D}$	1998	4638

* Numbers in this column represent $|\mathbb{S}|$ for the appropriate set $\mathbb{S}$.

** Numbers in this column represent $\sum_{g \in \mathbb{S}} H_g$ for the appropriate set $\mathbb{S}$.

3.4 Relevant data collection items

In general, we will use the letter g to index sites. Let H_g be the number of observed horizons for site g and let I_g be the inch at which the observed profile ends. For $g \in \mathcal{F} \cup \mathcal{L}$, I_g will usually be 48 inches, although some exceptions exist in the data. For $g \in \mathcal{S}$, I_g will be in the range of 6 to 12 inches. The value of H_g is limited to a maximum of 10 horizons by the data collection protocol. For $g \in \mathcal{S}$, H_g is 1 or 2. In general, we use $h = 1, \ldots, H_g$ to index horizons and $i = 1, \ldots, I_g$ to index inches for site g.

In the pilot project, soil texture is a horizon specific measurement. Horizon-based data present challenges in combining data across sites because the types and depths of horizons differ from site to site. This feature of the data structure is handled differently in the two analysis approaches. In both approaches, we wish to make use of all available texture data. This includes laboratory texture profiles, field texture profiles and surface horizon field texture measurements.

3.4.1 Horizon characteristics

For each observed horizon, the data collector records the horizon name. The horizon name includes the master horizon designation, denoted by a capital letter. The horizon name identifies general characteristics of the horizon, such as color, clay content or structure and special characteristics of the horizon, such as whether it is a buried horizon, a transitional horizon, etc.

For the data analyzed in Chapters 5 and 6, the possible values of master horizon designation are A, B and C. In general, A horizons are mineral horizons formed at or near the surface. B horizons form below A horizons. C horizons are layers that have not been strongly affected by soil-forming processes. Deviations may occur in this ordering (e.g., an A horizon may be buried below B and C horizons).

Table 3.2 shows an example of the horizon profile for one site. Note that a site may have multiple occurrences of A horizons and B horizons as demonstrated in this example. Multiple C horizons are also possible. Let m_{gh} and d_{gh} represent the master horizon designation and the horizon depth, respectively, of horizon h at site g. Horizon depth refers to the lower boundary of the horizon in inches. For example, if horizon h of site g occurs from inch 6 to 10, $d_{gh} = 10$.

Table 3.2 Example profile for site g.

Horizon index h	Horizon name	Master horizon m_{gh}	Horizon depth d_{gh}	Inch index i
1	Ap	A	7	$1, 2, \ldots, 7$
2	A	A	15	$8, \ldots, 15$
3	AB	A	24	$16, \ldots, 24$
4	Bg1	B	30	$25, \ldots, 30$
5	Bg2	B	36	$31, \ldots, 36$
6	BCg	B	42	$37, \ldots, 42$
7†	Cg	C	48††	$43, \ldots, 48$

† This is the value of H_g.

†† This is the value of I_g.

3.4.2 Soil texture variables

Soil texture is described by the proportions of clay, sand and silt which are present. Let c_1, c_2, and c_3 represent the proportions of clay, sand and silt, respectively, in the soil. These three proportions must sum to 1.0. This type of data is often called compositional data. We use the term *texture* to mean the vector $\mathbf{c} = (c_1, c_2, c_3)$. In this study, both clay and sand content are recorded for each horizon. Silt content is calculated by subtraction. Clay, sand and silt are called the components of soil texture. Figure 3.1 shows a ternary diagram for soil texture. This diagram is known as the texture triangle. It is a method of displaying all possible values of texture. The different sections of the triangle represent different texture classes.

Due to the multi-phase sampling design used for data collection, the amount of soil texture data collected varies from site to site. For sites in $\mathcal{S}$, we have a texture description based on determinations made in the field for the surface horizons only. This may include one or two horizons. For sites in $\mathcal{F} \cup \mathcal{L}$, we have field texture profiles for all horizons to a depth of 48 inches. For sites in $\mathcal{L}$, we also have laboratory profiles of texture for all horizons to a depth of 48 inches.

We will use superscripts to denote field and laboratory measurements of texture. Let the field determination for horizon h of site g be denoted $\left(c^{(f)}_{gh,1}, c^{(f)}_{gh,2}, c^{(f)}_{gh,3}\right)$ and the laboratory determination for the same horizon be $\left(c^{(l)}_{gh,1}, c^{(l)}_{gh,2}, c^{(l)}_{gh,3}\right)$. The field and laboratory measurements are both estimates of the true soil texture. Investigating the distribution of true soil texture is difficult because the variance of the measurement error in laboratory and field measurements is unknown. We suspect that this error is a significant source of variability in both types of measurements. However, it is expected that the laboratory measurements are more objective than the field measurements.

3.5 Transformation

Soil texture is a three-dimensional vector which must lie in a two-dimensional space because of the constraint on the sum of the components. A representation of this two-

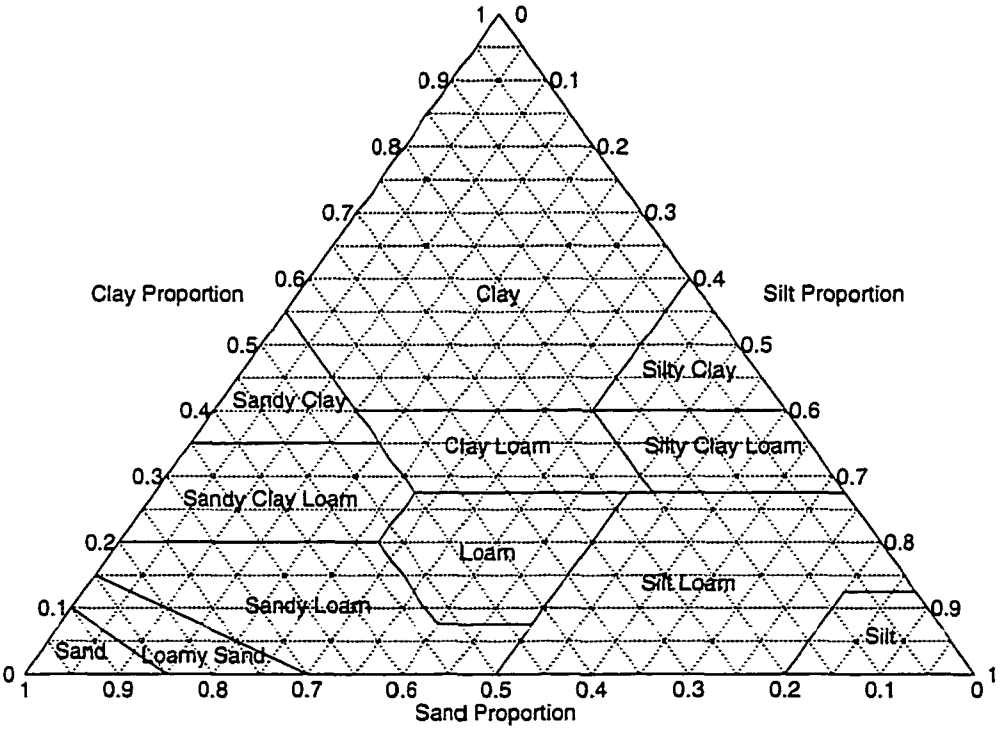

Figure 3.1 The texture triangle is a ternary diagram of soil texture values. The horizontal axis represents the proportion of sand; the 60° axis represents the clay proportion; and the −60° axis represents the silt proportion. The labeled sections represent texture classes.

dimensional region is presented in Figure 3.1. A transformation of texture will be used to create a two-dimensional vector with components which are not necessarily independent, but can take on any value in $\mathbb{R}^2$, the two-dimensional real number space. The estimation procedures presented in Chapters 5 and 6 are designed to analyze two-dimensional variables with this property.

Compositional data are often analyzed by using a log-ratio transformation as described in Aitchison (1986). A log-ratio transformation of a texture, (c_1, c_2, c_3), is defined by the function $\mathbb{L}$, where

$$\mathbb{L}(c_1, c_2, c_3) = \left(\log\left(\frac{c_1}{c_3}\right), \log\left(\frac{c_2}{c_3}\right)\right), \tag{3.1}$$

for $c_1 > 0$, $c_2 > 0$ and $c_3 > 0$. The log-ratio transformation maps a three-dimensional vector of positive values to $\mathbb{R}^2$. This transformation was chosen because it is documented in the literature and because standard methodology exists for analyzing data in $\mathbb{R}^2$.

The soil texture data contains some observations where one or more texture components are zero. The transformation $\mathbb{L}$ is not defined for any texture vector which contains a component equal to zero. To remedy this, an adjustment vector of small non-negative numbers is added to each texture observation before applying the transformation to eliminate zeroes within the data set. This shifts the entire dataset away from $(0, 0, 0)$, but should not significantly affect the overall analysis. Note that adding such a vector to each observation will not maintain the sum constraint for each vector. However, the transformation, $\mathbb{L}$, does not require the sum constraint.

For the data described in Section 3.3, sand is the only component for which some observations are zero. The adjustment vector was chosen by considering scatter plots of transformed variables using the following considerations. For an adjustment vector with components too close to zero, the transformation, $\mathbb{L}$, is defined, but the transformed values are very far from the rest of the data points. These outlying points may have too much influence when fitting regression models and may not reasonably reflect the relationships within the data. The adjustment vector was selected to prevent the transformed values of texture observations with components equal to zero from exhibiting this behavior. The vector $(0.005, 0.02, 0.0)$ is added to each observation.

Transformed laboratory and field textures are denoted $\boldsymbol{l}$ and $\boldsymbol{f}$, respectively. The bold-faced notation reflects the fact that these are vectors. That is, for horizon h of site g, we have

$$\begin{aligned} \boldsymbol{l}_{gh} &= (l_{gh,1}, l_{gh,2}) \\ &= \mathbb{L}\left(c^{(l)}_{gh,1} + 0.005, c^{(l)}_{gh,2} + 0.02, c^{(l)}_{gh,3} + 0.0\right) \\ \text{and} \quad \boldsymbol{f}_{gh} &= (f_{gh,1}, f_{gh,2}) \\ &= \mathbb{L}\left(c^{(f)}_{gh,1} + 0.005, c^{(f)}_{gh,2} + 0.02, c^{(f)}_{gh,3} + 0.0\right). \end{aligned}$$

For any vector $(x_1, x_2) \in \mathbb{R}^2$, the inverse of the log-ratio transformation is

$$\begin{aligned} (\hat{c}_1, \hat{c}_2, \hat{c}_3) &= \mathbb{L}^{-1}(x_1, x_2) - (0.005, 0.02, 0.0) \\ &= (\exp x_1 + \exp x_2 + 1)^{-1} (\exp x_1, \exp x_2, 1) (0.005, 0.02, 0.0). \end{aligned} \tag{3.2}$$

By construction, $\mathbb{L}^{-1}$ creates a vector that satisfies the sum constraint. However, we subtract the adjustment vector to completely back-transform the data.

4 QUANTILE ESTIMATION INCORPORATING AUXILIARY INFORMATION

4.1 Introduction

In survey sampling, we are often interested in studying the distribution of a particular variable of interest, Y. A convenient way to summarize a distribution function is by estimating quantiles of the distribution. By the pth quantile of a distribution, we mean the value Q such that $\mathbb{P}(Y \leq Q) = p$. Naturally, we are also usually interested in the quality of this estimate, so a method of variance estimation is desired.

One method of developing quantile estimators is to invert a distribution function estimator. Let $\hat{F}(t)$ denote an estimator of $\mathbb{P}(Y \leq t)$. Because the estimator $\hat{F}$ is often a step function, the form of the quantile estimator may not be smooth. The discontinuity may make variance estimation difficult. In particular, the methods of linearization and jackknifing may not be applicable.

In this chapter, we investigate a quantile estimator derived from a model-based distribution function estimator which incorporates auxiliary information. This distribution function estimator was introduced by Chambers and Dunstan (1986), so we refer to it as the CD estimator, or CDE. The quantile estimator is based on inverting the CDE. We derive a Bahadur representation for the quantile estimator. This representation can be used to derive an analytic expression for the asymptotic variance of the estimator. A simulation study is presented to evaluate the small-sample performance of the asymptotic variance expression.

4.2 Previous work

4.2.1 Notation

Consider a variable of interest, Y, and its distribution function $F_Y(y)$. Let $F'_Y(y)$ be the derivative of F_Y. We define the pth population quantile by

$$Q(p) = \inf\{q : F_Y(q) \geq p\}.$$

We are interested in estimating $Q(p)$ and in calculating standard errors to accompany the estimate. Consider a sequence of finite populations, $U_N = \{1, \ldots, i, \ldots, N\}$, of size N as $N \to \infty$. Let $y_1, \ldots, y_N$ denote the values of Y for each element in the finite population. We will use the notation F_N to denote the finite population distribution function, which is defined as

$$F_N(t) = N^{-1} \sum_{i \in U_N} \mathbb{I}(y_i \leq t), \tag{4.1}$$

where

$$\mathbb{I}(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

Then the pth quantile of the finite population is defined as

$$Q_N(p) = \inf\{q : F_N(q) \geq p\}. \tag{4.2}$$

We obtain data by taking a random sample, s_n, of size n from U_N using a sampling design $\mathcal{P}(\cdot)$. A point estimate of $Q(p)$, $Q_n(p)$, may be calculated by inverting the weighted empirical distribution function,

$$F_n(t) = \sum_{j \in s_n} w_j \mathbb{I}(y_j \leq t). \tag{4.3}$$

The customary design-based estimator of F_N uses $w_j = \pi_j \left(\sum_{i \in s_n} \pi_i\right)^{-2}$, where $\pi_j = \mathbb{P}(j \in s_n)$ is the inclusion probability of element $j \in U_N$.

4.2.2 Order statistics

Let $Y_{(r)}$ denote the rth order statistic of s_n. That is, there are exactly r observations in s_n that are less than or equal to $Y_{(r)}$. If, in (4.3), $w_j = n^{-1}$ for all $j \in s_n$, then $Q_n(p) = Y_{(r)}$ for $p \in \left(n^{-1}(r-1), n^{-1}r\right]$.

If F'_Y exists in a neighborhood of $Q(p)$, then for $p = r/n$, $Y_{(r)}$ is asymptotically normal in the sense that

$$\sqrt{n}(Y_{(r)} - Q(p)) \rightarrow_d \mathbb{N}\left(Q(p), \frac{p(1-p)}{[F'_Y(Q(p))]^2}\right)$$

as $n \rightarrow \infty$ (Serfling, 1980, p.77), where$\rightarrow_d$ denotes convergence in distribution and $\mathbb{N}(\cdot,\cdot)$ denotes a normal distribution. An estimate of the variance of this asymptotic distribution can be used to quantify a researcher's confidence in $Y_{(r)}$ as an estimate of $Q(p)$.

Another method of quantifying the researcher's confidence in an estimate is to use confidence intervals. If F_Y is a continuous function, then the random interval $(Y_{(r)}, Y_{(r')})$ contains $Q(p)$ with probability

$$\sum_{i=r}^{r'-1} \binom{n}{i} p^i (1-p)^{n-i}. \tag{4.4}$$

If F_Y is not continuous, Equation (4.4) is a lower bound for the confidence coefficient of the corresponding closed interval $[Y_{(r)}, Y_{(r')}]$ (David, 1981, p.15-16). An asymptotic confidence coefficient (instead of just a lower bound) for the closed interval can be obtained using normal approximations.

McCarthy (1965) investigates symmetric confidence intervals of the form

$$[Y_{(r)}, Y_{(n-r+1)}].$$

He gives conditions under which (4.4) is the exact confidence coefficient for a symmetric confidence interval for the median when the data is collected under a stratified sampling design. However, for the case of a discontinuous F_Y, Meyer (1972) gives a counter-example to McCarthy's result.

The above confidence interval methods allow the researcher a limited list of possible confidence coefficients for intervals. Woodruff (1952) presents a method for constructing approximate confidence intervals for a finite population quantile for any desired confidence coefficient. Woodruff's method relies on using a normal approximation to obtain an approximate $(1-\alpha)\%$ confidence interval for $F_n(Q_n(p))$. The endpoints of this interval are then inverted through F_n to provide endpoints for a confidence interval for $Q(p)$. The confidence coefficient of the resulting interval is taken to be approximately $(1-\alpha)\%$.

Sedransk and Meyer (1978) recommend a method for confidence intervals for quantiles when simple random sampling or stratified sampling is used. They recommend a method which ignores any stratification that was used at the design stage. They provide exact confidence coefficients for confidence intervals of the form $[Y_{(r)}, Y_{(r')}]$. However, the calculations can be cumbersome. Smith and Sedransk (1983) develop approximations to a lower bound for Sedransk and Meyer's confidence coefficients which are computationally simpler.

Francisco and Fuller (1990) consider estimation of quantiles for complex sampling designs. A test-inversion confidence set is presented. They show that construction of this set requires fewer conditions than those needed for Woodruff's method. However, the test-inversion confidence set may contain disjoint subsets and under some conditions, the two procedures are asymptotically equivalent.

4.2.3 Estimation using auxiliary information

In many surveys, auxiliary information is available. This information may take the form of a variable, X, the value of which is known for all elements of U_N or for a larger sample than s_n. Alternatively, only summary information may be available, e.g., means, totals or histogram information for X. If X is related to Y, we might use this information to improve upon estimators which do not incorporate auxiliary information. We review distribution function estimators which incorporate auxiliary information. Quantile estimators can be derived from these estimators.

4.2.3.1 Distribution function estimation assuming a linear model

Chambers and Dunstan (1986) assume that X is observed for each element in U_N and that Y and X are linearly related. They present a model-based estimator (CDE) which offers significant gains over the sample distribution function if the model is correct. Residuals from a fitted model are used to predict the value of an indicator for non-sampled elements.

Dunstan and Chambers (1989) extend the results of Chambers and Dunstan (1986) to the case of limited auxiliary information. That is, only histogram summaries of X are available. The estimator performs almost as well as the original CDE.

Rao, Kovar and Mantel (1990) propose a design-based estimator under the same assumptions as Chambers and Dunstan. Sampling weights are incorporated into the estimator to overcome design bias. A difference estimator is used to predict the sum of indicators for the non-sampled elements of U_N. The recommended estimator (RKME) is asymptotically design-unbiased and model-unbiased, but requires much more computation than the Chambers and Dunstan estimator.

Dorfman (1993) suggests a modification to the RKME which does not require second-order inclusion probabilities. Dorfman concludes that the modified estimator is preferred when a linear model is fit without thorough consideration of goodness of fit. However, if the linear model fits well, a model-based estimator such as the CDE performs better.

Chambers, Dorfman and Hall (1992) further investigate the asymptotic properties of the CDE and the RKME. Asymptotic expressions for the variance of each are compared. Simulation studies of populations for which the model is correct appear to indicate that the CDE has smaller variance than the RKME. Cases can be constructed where this is not true. However, the authors suggest that these cases do not often occur in practice.

Wang and Dorfman (1996) consider combining the CDE and the RKME. They propose using a weighted average of the two estimators, where the weight is estimated to minimize the asymptotic mean squared error the resulting estimator. Goyeneche (1999) investigates an extension of the CDE called the local residuals estimator. The CDE is constructed under the model that the residuals are homoskedastic or have known vari-

ance function. If neither of these is the case, but the variance of the residuals changes smoothly, Goyeneche's local residuals estimator may be appropriate.

4.2.3.2 Nonparametric superpopulation model

The linear superpopulation model assumed for the work mentioned in Section 4.2.3.1 can be relaxed. A nonparametric superpopulation model with less restrictive assumptions about the form of the joint distribution of X and Y can be considered.

Kuo (1988) proposes two estimators: a kernel estimator and a k nearest neighbor estimator. These are regression type estimators developed by fitting a non-parametric regression to X and an indicator function of Y. Chambers, Dorfman and Wehrly (1993) point out that Kuo's kernel estimator is biased if the linear model assumed by Chambers and Dunstan is true. They suggest a bias calibration for the estimator which should perform similarly to Kuo's estimator if the population is not linear and should be more efficient if the population is approximately linear.

In Kuk (1993), the conditional distribution of Y given X in the finite population is estimated by a smoothed nonparametric method. The estimate of F_Y is obtained by averaging the estimates of the conditional distribution over the observed values of X. A drawback of ths method is that it requires a smoothed, preferably nonparametric, estimate of the joint distribution of X and Y.

Dorfman and Hall (1993) provide large sample theory for six nonparametric estimators of the distribution function. One of the estimators was that suggested by Kuo (1988). Two are modifications of the CDE and RKME which use predictions and fitted residuals from a non-parametric regression. Another is the nonparametric calibration estimator presented by Chambers, Dorfman and Wehrly (1993).

4.2.4 Chambers and Dunstan distribution function estimator

We consider a simplified version of the model assumed in Chambers and Dunstan (1986) and later works. Assume that Y and X follow the superpopulation model

$$Y = X\beta + E \tag{4.5}$$

where $X \sim F_X$ and is independent of $E \sim F_E$. Note that

$$F_Y(q) = \int \int_{-\infty}^{q-x\beta} dF_E(e)dF_X(x) = \int F_E(q - \beta x)dF_X(x).$$

As before, $y_1, \ldots, y_N$ denote the values of Y in U_N. Also, let $x_1, \ldots, x_N$ and $e_1, \ldots, e_N$ denote the values of X and E, respectively, in U_N. Suppose the value of Y is only observed for elements of s_n, while the values of X are available for every element in U_N.

Let $u_n = \sqrt{n}(\hat{\beta} - \beta)$, where $\hat{\beta}$ is the ordinary least squares estimator of β. As an estimator of (4.1), Chambers and Dunstan proposed

$$\hat{F}_N(q; u_n) = N^{-1}\left[\sum_{j \in s_n} \mathbb{I}(y_j \leq q) + \sum_{i \notin s_n} n^{-1} \sum_{j \in s_n} \mathbb{I}(\hat{y}_{ij}(u_n) \leq q)\right], \tag{4.6}$$

where $\hat{y}_{ij}(u_n) = \left(\beta + \frac{u_n}{\sqrt{n}}\right) x_i + \hat{e}_j(u_n)$ and $\hat{e}_j(u_n) = y_j - \left(\beta + \frac{u_n}{\sqrt{n}}\right) x_j$. We refer to (4.6) as the CDE. The CDE has the desirable property that if $Y \propto X$, the CDE is exactly $F_N(q)$. However, in general, it is not design-unbiased.

The estimator (4.6) can be viewed as a weighted sum of two estimators of F_N: one from observed values and one from the auxiliary information of the out-of-sample elements. That is,

$$\hat{F}_N(q; u_n) = N^{-1}\left[nF_n(q) + (N - n)F_r^*(q; u_n)\right],$$

where $F_n(\cdot)$ is the usual emprical distribution function and

$$F_r^*(q; u_n) = n^{-1} \sum_{j \in s_n} (N - n)^{-1} \sum_{i \notin s_n} \mathbb{I}(\hat{y}_{ij}(u_n) \leq q).$$

The CDE also has the convenient interpretation of being a weighted empirical distribution function of the data after imputation. The imputed values are $\{\hat{y}_{ij}(u_n)\}$. Each original observation is given a weight, $w_j = 1$. The imputed values are each given a weight of $w_{ij} = n^{-1}$.

Based on some regularity conditions, Chambers and Dunstan show that as both n and N increase,

$$\left[\left(1 - \frac{n}{N}\right)^2 \{W_r^*(q, \beta) + W_r(q, \beta)\}\right]^{-\frac{1}{2}} \{\hat{F}_N(q; u_n) - F_N(q)\}$$

converges in distribution to a standard normal random variable, where

$$W_r(q,\beta) = (N-n)^{-2} \sum_{i \in U_N} F_E(q - x_i\beta)(1 - F_E(q - x_i\beta))$$

and

$$W_r^*(q,\beta) = \boldsymbol{D}_r(q,\beta)\boldsymbol{V}_r^*(q,\beta)\boldsymbol{D}_r'(q,\beta),$$

where $\boldsymbol{V}_r^*(q,\beta)$ is the covariance matrix of $(F_r^*(q;0) - \mathbb{E}(F_r^*(q;0)), \hat{\beta} - \beta)$ and

$$\boldsymbol{D}_r(q,\beta) = \left(1, n^{-1}\sum_{j \in s_n}(N-n)^{-1}\sum_{i \notin s_n}(x_j - x_i)F_E'(q - x_i\beta)\right).$$

However, Chambers and Dunstan do not provide the form of $\boldsymbol{V}_r^*(q,\beta)$.

Chambers, Dorfman and Hall (1992) further investigated the variance of the CDE and another distribution function estimator. They derived an analytic expression for the variance of $\hat{F}_N(q;u) - F_N(q)$. Let F_E' denote the derivative of F_E. Define

$$\begin{aligned} \lambda &= \lim_{N\to\infty} nN^{-1} \\ \text{and } \Gamma &= \mathrm{Cov}(X, F_E'(q - X\beta)). \end{aligned}$$

Then

$$\begin{aligned} \mathrm{Var}(\hat{F}_N(q;u_n) - F_N(q)) = n^{-1}(1-\lambda)^2\Big\{ & N^{-1}(1-\lambda)(F_Y(q) - F_Y^2(q)) + n\mathrm{Var}(\hat{\beta})\Gamma^2 \\ & + \int\int F_E((q - x_1\beta) \wedge (q - x_2\beta))dF_X(x_1)dF_X(x_2) - F_Y^2(q)\Big\} + o(n^{-1}), \end{aligned} \tag{4.7}$$

where $a \wedge b$ denotes the minimum of a and b. This expression is consistent with the result in the Chambers and Dunstan paper, but provides more detail of the form of $\boldsymbol{V}_r^*(q,\beta)$.

4.2.5 Quantile estimator derived from CDE

Chambers and Dunstan recognize that a common use of distribution function estimators is to provide quantile estimators. Define an estimator for the pth quantile as

$$\hat{Q}_N(p;u_n) = \inf\{q : \hat{F}_N(q;u_n) \geq p\}. \tag{4.8}$$

This is a common way to define a quantile estimator based on a distribution function estimator. Chambers and Dunstan state that this estimator is asymptotically unbiased for the finite population quantile,

$$Q_N(p) = \inf\{q : F_N(q) \geq p\}.$$

They note that the finite population quantile has the Bahadur representation

$$Q_N(p) = Q(p) + \frac{p - F_N(Q(p))}{F'_Y(Q(p))} + o_p(N^{-\frac{1}{2}}),$$

where $Q(p)$ is the pth quantile of F_Y. They also conjecture that (4.8) has the similar representation

$$\hat{Q}_N(p; u_n) = Q(p) + \frac{p - \hat{F}_N(Q(p); u_n)}{F'_Y(Q(p))} + o_p(n^{-\frac{1}{2}}).$$

If the conjecture is valid, then according to the Chambers, Dorfman and Hall results, the asymptotic variance of $(\hat{Q}_N(p; u_n) - Q_N(p))$ is

$$\left(1 - \frac{n}{N}\right)^2 [F'_Y(Q(p))]^{-2} [W_r^*(Q(p), \beta) + W_r(Q(p), \beta)].$$

In the next section, we investigate this conjecture.

4.3 Properties of the Chambers and Dunstan quantile estimator

A linearized form of the Chambers and Dunstan quantile estimator $\hat{Q}_N(p; u_n)$ is desired in order to obtain an asymptotic variance expression. Recall that $Q_n(p)$ is the pth sample quantile. Bahadur (1966) shows that if F_Y has two derivatives in a neighborhood of $Q(p)$, the second derivative is bounded in this neighborhood and $F'_Y(Q(p))$ is positive, then

$$Q_n(p) = Q(p) + \frac{p - F_n(Q(p))}{F'_Y(Q(p))} + O\left(n^{-\frac{3}{4}} \log n\right).$$

Ghosh (1971) provides a weaker version of this result with fewer assumptions. Define

$$Q_L(p_n) = Q(p) + \frac{p_n - p}{F'_Y(Q(p))}.$$

If $F_Y'(Q(p))$ exists and is strictly positive and $p_n - p = O\left(n^{-\frac{1}{2}}\right)$, then

$$Q_n(p_n) = Q_L(p_n) + \frac{p - F_n(Q(p))}{F_Y'(Q(p))} + o_p\left(n^{-\frac{1}{2}}\right).$$

For a fixed value of u, we will develop a similar representation for $\hat{Q}_N(p;u)$ which also includes a term which is a linear function of u. Section 4.3.1 contains a proof of a Bahadur-type representation for the estimator. An asymptotic variance expression is given in section 4.3.2.

4.3.1 Bahadur representation for $\hat{Q}_N$ for fixed u

We wish to show that the estimator given in (4.8) has a Bahadur type representation. Define the sequence β_N^+

$$\beta_N^+ = \beta + \frac{u}{\sqrt{n}}$$

for a fixed number u. A Bahadur representation is presented in Theorem 4.1 for a fixed value of u. The following lemmas will be useful in the proof of Theorem 4.1.

Lemma 4.1 *If we have that*

C1. *for all t, $F_Y'(t) \leq M$ for some $M > 0$*

C2. *for all t, $F_Y''(t)$ exists and*

C3.

$$\begin{aligned}
&\int\int F_Y'\left(Q(p) - \left(\beta + \frac{u}{\sqrt{n}}\right)(z-x)\right) dF_X(x)dF_X(z) \\
&= \int\int F_Y'\left(Q(p) - \beta(z-x)\right)dF_X(x)dF_X(z) \\
&\quad - \frac{u}{\sqrt{n}}\int\int (z-x)F_Y''\left(Q(p) - \beta(z-x)\right)dF_X(x)dF_X(z) + o\left(n^{-\frac{1}{2}}\right),
\end{aligned} \tag{4.9}$$

then

$$\int\int F_Y'\left(Q(p) - \left(\beta + \frac{u}{\sqrt{n}}\right)(z-x)\right)dF_X(x)dF_X(z) = F_Y'\left(Q(p)\right) + \frac{u}{\sqrt{n}}\kappa + o\left(n^{-\frac{1}{2}}\right),$$

where

$$\kappa = -\int\int (z-x)F_Y''\left(Q(p)-\beta(z-x)\right)dF_X(x)dF_X(z)$$

Proof. The first term of the right hand side of (4.9) is

$$\begin{aligned}\int\int F_Y'\left(Q(p)-\beta(z-x)\right)&dF_X(x)dF_X(z)\\ &= \int\int \lim_{N\to\infty} N\left[F_Y\left(Q(p)-\beta(z-x)+\frac{1}{N}\right)\right.\\ &\qquad \left. - F_Y\left(Q(p)-\beta(z-x)\right)\right]dF_X(x)dF_X(z).\end{aligned} \tag{4.10}$$

Define

$$g_N(x,z) = N\left[F_Y\left(Q(p)-\beta(z-x)+\frac{1}{N}\right) - F_Y\left(Q(p)-\beta(z-x)\right)\right].$$

If the $\{g_N\}$ are uniformly integrable, then the integral and limit in (4.10) can be interchanged (Billingsley, 1995, p. 217). By the mean value theorem, we have that

$$g_N(x,z) = F_Y'\left(\xi_N(x,z)\right)$$

for some ξ_N in $\left(Q(p)-\beta(z-x), Q(p)-\beta(z-x)+\frac{1}{N}\right)$. Then since F_Y' is bounded by M, we have

$$\begin{aligned}&\lim_{\alpha\to\infty}\sup_N \int\int_{\{|g_N|\geq\alpha\}} |g_N| dF_X(x)dF_X(z)\\ &= \lim_{\alpha\to\infty}\sup_N \int\int_{\{|F_Y'(\xi_N)|\geq\alpha\}} \left|F_Y'\left(\xi_N\right)\right| dF_X(x)dF_X(z)\\ &\leq \lim_{\alpha\to\infty}\sup_N M\,\mathbb{I}\left(M\geq\alpha\right)\\ &= 0.\end{aligned}$$

Thus, the $\{g_N\}$ are uniformly integrable and we can interchange the limit and integral in (4.10) to arrive at

$$\lim_{N\to\infty} N\left[\int\int F_Y\left(Q(p)-\beta(z-x)+\frac{1}{N}\right) - F_Y\left(Q(p)-\beta(z-x)\right)dF_X(x)dF_X(z)\right]$$

$$
\begin{aligned}
&= \frac{d}{dt}\left[\int\int F_Y(t-\beta(z-x))dF_X(x)dF_X(z)\right]\Big|_{t=Q(p)} \\
&= \frac{d}{dt}\left[\mathbb{E}\left(\mathbb{P}\left(Y+\beta(z-x)\le t \,\middle|\, Z=z, X=x\right)\right)\right]\Big|_{t=Q(p)} \\
&= \frac{d}{dt}\left[\mathbb{E}\left(\mathbb{P}\left(\beta x+E+\beta(z-x)\le t \,\middle|\, Z=z, X=x\right)\right)\right]\Big|_{t=Q(p)} \\
&= \frac{d}{dt}\left[\mathbb{P}\left(\beta Z+E\le t\right)\right]\Big|_{t=Q(p)} \\
&= \frac{d}{dt}F_Y(t)\Big|_{t=Q(p)} \\
&= F_Y'\left(Q(p)\right).
\end{aligned}
$$

Thus, we have the result. ▲

Lemma 4.2 *If $\delta_N = o(1)$, then for all k,*

$$\mathbb{I}\left(0\le k\right) - \mathbb{I}\left(\delta_N \le k\right) = o(1)$$

Proof.

$$
\begin{aligned}
\lim_{N\to\infty} \mathbb{I}\left(0\le k\right) - \mathbb{I}\left(\delta_N \le k\right) &= \lim_{N\to\infty} \mathbb{I}\left(0\le k<\delta_N\right) - \mathbb{I}\left(\delta_N \le k<0\right) \\
&= 0,
\end{aligned}
$$

since k is fixed and $\delta_N \to 0$. ▲

Next, we present a theorem providing a Bahadur representation for a fixed value of u. The proof follows that of Ghosh (1971). Recall that

$$Q_L(p_N) = Q(p) + \frac{p_N - p}{F_Y'\left(Q(p)\right)}.$$

Theorem 4.1 *Suppose model (4.5) holds, we have conditions* C1, C2 *and* C3 *from Lemma 4.1 and*

C4. $\lim_{N\to\infty} nN^{-1} = \lambda > 0.$

C5. $(p_N - p) = O(n^{\frac{1}{2}}N^{-1}),$

then

$$\hat{Q}_N(p;u) = Q_L(p_N) + \frac{p - \hat{F}_N(Q(p);u)}{F_Y'\left(Q(p)\right)} + R_N(u), \tag{4.11}$$

where $R_N(u) = o_p\left(n^{-\frac{1}{2}}\right)$.

Proof. Define

$$V_N(u) = \frac{N}{\sqrt{n}} \left(\hat{Q}_N(p;u) - Q_L(p_N) \right). \tag{4.12}$$

Then for any real number k,

$$\begin{aligned} \{V_N(u) \le k\} &= \left\{ p_N \le \hat{F}_N \left(Q_L(p_N) + k\frac{\sqrt{n}}{N}; u \right) \right\} \\ &= \left\{ \frac{N}{\sqrt{n}} \left[\frac{F_Y \left(Q_L(p_N) + k\frac{\sqrt{n}}{N} \right) - \hat{F}_N \left(Q_L(p_N) + k\frac{\sqrt{n}}{N}; u \right)}{F_Y'(Q(p))} \right] \le k_N \right\} \end{aligned} \tag{4.13}$$

where

$$k_N = \frac{N}{\sqrt{n}} \left[\frac{F_Y \left(Q_L(p_N) + k\frac{\sqrt{n}}{N} \right) - p_N}{F_Y'(Q(p))} \right]. \tag{4.14}$$

Now

$$\begin{aligned} k_N &= \frac{N}{\sqrt{n}} [F_Y'(Q(p))]^{-1} \left[p + F_Y'(Q(p)) \left(\frac{p_N - p}{F_Y'(Q(p))} + k\frac{\sqrt{n}}{N} \right) + o(n^{\frac{1}{2}} N^{-1}) - p_N \right] \\ &= k + o(1) \to k \end{aligned} \tag{4.15}$$

as $N \to \infty$. Define

$$\begin{aligned} \delta_N &= k\frac{\sqrt{n}}{N} + \frac{p_N - p}{F_Y'(Q(p))} \\ Q_L^+(p_N) &= Q(p) + \delta_N \\ Z_N(q;u) &= \frac{N}{\sqrt{n}} [F_Y'(Q(p))]^{-1} \left(F_Y(q) - \hat{F}_N(q;u) \right). \end{aligned}$$

Then (4.13) and (4.15) yield that for all k and all $\varepsilon > 0$,

$$\begin{aligned} &\lim_{N\to\infty} \mathbb{P} \left(V_N(u) \le k, Z_N(Q_L^+(p_N);u) > k + \varepsilon \right) = 0 \\ &\lim_{N\to\infty} \mathbb{P} \left(V_N(u) > k + \varepsilon, Z_N(Q_L^+(p_N);u) \le k \right) = 0. \end{aligned} \tag{4.16}$$

We wish to obtain statements similar to (4.16) with $Z_N(Q_L^+(p_N);u)$ replaced by $Z_N(Q(p);u)$. In order to do this, we will show that

$$\mathbb{E} \left(Z_N(Q_L^+(p_N);u) - Z_N(Q(p);u) \right)^2 \to 0$$

and thus

$$Z_N(Q_L^+(p_N);u) - Z_N(Q(p);u) \to_p 0. \tag{4.17}$$

We will need an expression for the following expectation.

$$\begin{aligned}
&\frac{N^2}{n}\mathbb{E}\left(\hat{F}_N(q_1;u)\hat{F}_N(q_2;u)\right)\\
&=n^{-1}\mathrm{E}\left[\left(\sum_{j\in s_n}\mathrm{I}\{y_j\le q_1\}\sum_{j'\in s_n}\mathrm{I}\{y_{j'}\le q_2\}\right)\right.\\
&\quad+\left(\sum_{i\notin s_n}n^{-1}\sum_{j\in s_n}\mathbb{I}\left(\hat{y}_{ij}(u)\le q_1\right)\sum_{j'\in s_n}\mathbb{I}\left(y_{j'}\le q_2\right)\right)\\
&\quad+\left(\sum_{j\in s_n}\mathbb{I}\left(y_j\le q_1\right)\sum_{i'\notin s_n}n^{-1}\sum_{j'\in s_n}\mathbb{I}\left(\hat{y}_{i'j'}(u)\le q_2\right)\right)\\
&\quad+\left.\left(\sum_{i\notin s_n}n^{-1}\sum_{j\in s_n}\mathbb{I}\left(\hat{y}_{ij}(u)\le q_1\right)\sum_{i'\notin s_n}n^{-1}\sum_{j'\in s_n}\mathbb{I}\left(\hat{y}_{i'j'}(u)\le q_2\right)\right)\right]\\
&=F_Y\left(q_1\wedge q_2\right) &(4.18)\\
&\quad+(n-1)F_Y(q_1)F_Y(q_2) &(4.19)\\
&\quad+\frac{(N-n)}{n}\mathbb{P}\left(Y_j\le q_1,\beta_N^+X_{i'}+Y_j-\beta_N^+X_j\le q_2\right) &(4.20)\\
&\quad+\frac{(n-1)(N-n)}{n}F_Y(q_1)\mathbb{P}\left(\beta_N^+X_{i'}+Y_{j'}-\beta_N^+X_{j'}\le q_2\right) &(4.21)\\
&\quad+\frac{(N-n)}{n}\mathbb{P}\left(\beta_N^+X_i+Y_j-\beta_N^+X_j\le q_1,Y_j\le q_2\right) &(4.22)\\
&\quad+\frac{(n-1)(N-n)}{n}\mathbb{P}\left(\beta_N^+X_i+Y_j-\beta_N^+X_j\le q_1\right)F_Y(q_2) &(4.23)\\
&\quad+\frac{(N-n)}{n^2}\mathbb{P}\left(\beta_N^+X_i+Y_j-\beta_N^+X_j\le(q_1\wedge q_2)\right) &(4.24)\\
&\quad+\frac{(N-n)(n-1)}{n^2}\\
&\quad\times\mathbb{P}\left(\beta_N^+X_i+Y_j-\beta_N^+X_j\le q_1,\beta_N^+X_i+Y_{j'}-\beta_N^+X_{j'}\le q_2\right) &(4.25)\\
&\quad+\frac{(N-n)(N-n-1)}{n^2}\\
&\quad\times\mathbb{P}\left(\beta_N^+X_i+Y_j-\beta_N^+X_j\le q_1,\beta_N^+X_{i'}+Y_j-\beta_N^+X_j\le q_2\right) &(4.26)\\
&\quad+\frac{(N-n)(N-n-1)(n-1)}{n^2}\\
&\quad\times\mathbb{P}\left(\beta_N^+X_i+Y_j-\beta_N^+X_j\le q_1\right)\mathbb{P}\left(\beta_N^+X_i+Y_j-\beta_N^+X_j\le q_2\right), &(4.27)
\end{aligned}$$

where $a \wedge b$ denotes the minimum of a and b. Label the terms of the sum $T_1(q_1, q_2; u)$ through $T_{10}(q_1, q_2; u)$.

Now

$$\begin{aligned}
&\mathbb{E}\left(Z_N(Q_L^+(p_N); u) - Z_N(Q(p); u)\right)^2 = \\
&\left[F_Y'\left(Q(p)\right)\right]^{-2} \left\{ \sum_{i=1}^{10} T_i^*(u) - \frac{N^2}{n}\left[F_Y(Q_L^+(p_N)) - F_Y(Q(p))\right]^2 \right. \\
&\left. -\frac{2N^2}{n}\left[F_Y(Q_L^+(p_N)) - F_Y(Q(p))\right]\left[\mathbb{B}(Q_L^+(p_N); u) - \mathbb{B}(Q(p); u)\right] \right\},
\end{aligned} \tag{4.28}$$

where

$$\mathbb{B}(q; u) = \mathbb{E}\left(\hat{F}_N(q; u) - F_Y(q)\right)$$

and $T_i^*(u) = T_i(Q_L^+(p_N), Q_L^+(p_N); u) - 2T_i(Q_L^+(p_N), Q(p); u) + T_i(Q(p), Q(p); u)$.

We will now consider each of the T^* terms. Note that $T_7(q_1, q_2; u) = O\left(n^{-1}\right)$, $T_i(q_1, q_2; u), i \in \{1, 3, 5, 8, 9\}$ are $O(1)$ and $T_i(q_1, q_2; u), i \in \{2, 4, 6, 10\}$ are $O(n)$. Define $\delta_1 = q_1 - Q(p)$ and $\delta_2 = q_2 - Q(p)$. Note that δ_1 and δ_2 are either $\delta_N = O\left(n^{\frac{1}{2}} N^{-1}\right)$ or 0. Thus, the $O(n)$ terms will be expanded to second order terms, the $O(1)$ terms will be expanded to zeroth order terms, and the $O\left(n^{-1}\right)$ term will not be expanded at all.

Term (4.18) is

$$\begin{aligned}
T_1(q_1, q_2; u) &= F_Y(q_1)\mathbb{I}\left(q_1 - q_2 \leq 0\right) + F_Y(q_2)\mathbb{I}\left(q_1 - q_2 > 0\right) \\
&= \mathbb{I}\left(q_1 - q_2 \leq 0\right)\left[p + \mathbb{I}\left(\delta_1 \neq 0\right) o(1)\right] \\
&\quad + \mathbb{I}\left(q_1 - q_2 > 0\right)\left[p + o(1)\mathbb{I}\left(\delta_2 \neq 0\right)\right] \\
&= p + o(1)\left[\mathbb{I}\left(\delta_1 \neq 0\right) + \mathbb{I}\left(\delta_2 \neq 0\right)\right].
\end{aligned} \tag{4.29}$$

Thus $T_1^*(u) = o(1)$.

Term (4.19) is

$$\begin{aligned}
T_2(q_1, q_2; u) =& (n-1)\left[p + \delta_1 F_Y'\left(Q(p)\right) + \delta_1^2 \frac{F_Y''\left(Q(p)\right)}{2} + \mathbb{I}\left(\delta_1 \neq 0\right) o\left(\frac{n}{N^2}\right)\right] \\
&\times \left[p + \delta_2 F_Y'\left(Q(p)\right) + \delta_2^2 \frac{F_Y''\left(Q(p)\right)}{2} + \mathbb{I}\left(\delta_2 \neq 0\right) o\left(\frac{n}{N^2}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
=&(n-1)\left[p^2+(\delta_1+\delta_2)\,pF_Y'\left(Q(p)\right)\right.\\
&\left.+\left(\delta_1^2+\delta_2^2\right)p\frac{F_Y''\left(Q(p)\right)}{2}+\delta_1\delta_2\left[F_Y'\left(Q(p)\right)\right]^2\right]\\
&+o(1)\left[\mathbb{I}\left(\delta_1\neq 0\right)+\mathbb{I}\left(\delta_2\neq 0\right)\right]. \qquad (4.30)
\end{aligned}
$$

Thus

$$
\begin{aligned}
T_2^*(u) &= (n-1)\left\{p^2+2\delta_N pF_Y'\left(Q(p)\right)+\delta_N^2 pF_Y''\left(Q(p)\right)\right.\\
&\quad +\delta_N^2\left[F_Y'\left(Q(p)\right)\right]^2-2p^2-2\delta_N pF_Y'\left(Q(p)\right)\\
&\quad \left.-\delta_N^2 pF_Y''\left(Q(p)\right)+p^2\right\}+o(1)\\
&= (n-1)\delta_N^2\left[F_Y'\left(Q(p)\right)\right]^2+o(1). \qquad (4.31)
\end{aligned}
$$

Term (4.20) is

$$
\begin{aligned}
&T_3(q_1,q_2;u)=\frac{(N-n)}{n}\int\int F_Y(q_1\wedge q_2-\beta_N^+(z-x))dF_X(z)dF_X(x)\\
=&\frac{(N-n)}{n}\int\int\left[F_Y(q_1)\mathbb{I}\left(q_1-q_2\leq-\beta_N^+(z-x)\right)\right.\\
&\left.+F_Y(q_2-\beta_N^+(z-x))\mathbb{I}\left(q_1-q_2>-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right]\\
=&\frac{(N-n)}{n}\int\int\left\{\mathbb{I}\left(q_1-q_2\leq-\beta_N^+(z-x)\right)\left[p+o(1)\mathbb{I}\left(\delta_1\neq 0\right)\right]\right.\\
&\left.+\mathbb{I}\left(q_1-q_2>-\beta_N^+(z-x)\right)\left[F_Y\left(Q(p)-\beta_N^+(z-x)\right)+o(1)\mathbb{I}\left(\delta_2\neq 0\right)\right]\right\}\\
&\times dF_X(z)dF_X(x)\\
=&\frac{(N-n)}{n}\left\{p\int\int\mathbb{I}\left(q_1-q_2\leq-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right.\\
&\left.+\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(q_1-q_2>-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right\}\\
&+o(1)\left[\mathbb{I}\left(\delta_1\neq 0\right)+\mathbb{I}\left(\delta_2\neq 0\right)\right]. \qquad (4.32)
\end{aligned}
$$

Thus

$$
\begin{aligned}
T_3^*(u)=&\frac{(N-n)}{n}\left\{p\int\int\mathbb{I}\left(0\leq-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right.\\
&+\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(0>-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\\
&\left.-2p\int\int\mathbb{I}\left(\delta_N\leq-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right.
\end{aligned}
$$

$$
\begin{aligned}
&- 2\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(\delta_N > -\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\\
&+ p\int\int \mathbb{I}\left(0 \le -\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\\
&+ \int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(0 > -\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\Big\}\\
&+ o(1)\\
=&\frac{2(N-n)}{n}\left\{\int\int\left[\mathbb{I}\left(0 \le -\beta_N^+(z-x)\right) - \mathbb{I}\left(\delta_N \le -\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right]\right.\\
&\left.\times\left[p + F_Y\left(Q(p)-\beta_N^+(z-x)\right)\right]dF_X(z)dF_X(x)\right\} + o(1)\\
=&o(1),
\end{aligned}
\tag{4.33}
$$

using Lemma 4.2.

Term (4.21) is

$$
\begin{aligned}
&T_4(q_1,q_2) = \frac{(n-1)(N-n)}{n}F_Y(q_1)\int\int F_Y(q_2-\beta_N^+(z-x))dF_X(z)dF_X(x)\\
=&\frac{(n-1)(N-n)}{n}\left[p + \delta_1 F_Y'\left(Q(p)\right) + \delta_1^2\frac{F_Y''\left(Q(p)\right)}{2} + \mathbb{I}\left(\delta_1 \neq 0\right)o\left(\frac{n}{N^2}\right)\right]\\
&\times\int\int\left[F_Y\left(Q(p)-\beta_N^+(z-x)\right) + \delta_2 F_Y'\left(Q(p)-\beta_N^+(z-x)\right)\right.\\
&\left.+\delta_2^2\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2} + \mathbb{I}\left(\delta_2 \neq 0\right)o\left(\frac{n}{N^2}\right)\right]dF_X(z)dF_X(x)\\
=&\frac{(n-1)(N-n)}{n}\left\{p\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right.\\
&+ p\times\\
&\int\int\delta_2 F_Y'\left(Q(p)-\beta_N^+(z-x)\right) + \delta_2^2\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2}dF_X(z)dF_X(x)\\
&+ F_Y'\left(Q(p)\right)\times\\
&\int\int\delta_1 F_Y\left(Q(p)-\beta_N^+(z-x)\right) + \delta_1\delta_2 F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\\
&\left.+\delta_1^2\frac{F_Y''\left(Q(p)\right)}{2}\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right\}\\
&+ o(1)\left[\mathbb{I}\left(\delta_1 \neq 0\right) + \mathbb{I}\left(\delta_2 \neq 0\right)\right].
\end{aligned}
\tag{4.34}
$$

Thus

$$\begin{aligned}T_4^{*} =& \frac{(n-1)(N-n)}{n}\left\{\delta_N p\int\int F_Y'\left(Q(p)-\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\right.\\ &-\delta_N F_Y'\left(Q(p)\right)\int\int F_Y\left(Q(p)-\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\\ &+\delta_N^2 p\int\int\frac{F_Y''\left(Q(p)-\beta_N^{+}(z-x)\right)}{2}dF_X(z)dF_X(x)\\ &-\delta_N^2\frac{F_Y''\left(Q(p)\right)}{2}\int\int F_Y\left(Q(p)-\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\\ &\left.+\delta_N^2 F_Y'\left(Q(p)\right)\int\int F_Y'\left(Q(p)-\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\right\}+o(1).\end{aligned} \tag{4.35}$$

Term (4.22) is

$$\begin{aligned}T_5(q_1,q_2) &= \frac{(N-n)}{n}\int\int F_Y(q_1-\beta_N^{+}(z-x)\wedge q_2)dF_X(z)dF_X(x)\\ &= \frac{(N-n)}{n}\times\\ &\quad\left\{\int\int F_Y(q_1-\beta_N^{+}(z-x))\mathbb{I}\left(q_1-q_2\le\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\right.\\ &\quad\left.+\int\int F_Y(q_2)\mathbb{I}\left(q_1-q_2>\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\right\}\\ &= \frac{(N-n)}{n}\int\int\left\{F_Y\left(Q(p)-\beta_N^{+}(z-x)\right)\mathbb{I}\left(q_1-q_2\le\beta_N^{+}(z-x)\right)\right.\\ &\quad\left.+p\mathbb{I}\left(q_1-q_2>\beta_N^{+}(z-x)\right)\right\}dF_X(z)dF_X(x)\\ &\quad+o(1)\left[\mathbb{I}\left(\delta_1\neq 0\right)+\mathbb{I}\left(\delta_2\neq 0\right)\right].\end{aligned} \tag{4.36}$$

Thus

$$\begin{aligned}T_5^{*} =& \frac{(N-n)}{n}\left\{\int\int F_Y\left(Q(p)-\beta_N^{+}(z-x)\right)\mathbb{I}\left(0\le\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\right.\\ &+p\int\int\mathbb{I}\left(0>\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\\ &-2\int\int F_Y\left(Q(p)-\beta_N^{+}(z-x)\right)\mathbb{I}\left(\delta_N\le\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\\ &-2p\int\int\mathbb{I}\left(\delta_N>\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\\ &+\int\int F_Y\left(Q(p)-\beta_N^{+}(z-x)\right)\mathbb{I}\left(0\le\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\\ &\left.+p\int\int\mathbb{I}\left(0>\beta_N^{+}(z-x)\right)dF_X(z)dF_X(x)\right\}+o(1)\end{aligned}$$

$$
\begin{aligned}
=&\frac{2(N-n)}{n}\left\{\int\int\left[\mathbb{I}\left(0\leq\beta_N^+(z-x)\right)-\mathbb{I}\left(\delta_N\leq\beta_N^+(z-x)\right)\right]\right.\\
&\left.\times\left[p+F_Y\left(Q(p)-\beta_N^+(z-x)\right)\right]dF_X(z)dF_X(x)\right\}+o(1)\\
=&o(1),
\end{aligned}
\tag{4.37}
$$

using Lemma 4.2.

Term (4.23) is

$$
\begin{aligned}
T_6(q_1,q_2)=&\frac{(n-1)(N-n)}{n}F_Y(q_2)\int\int F_Y(q_1-\beta_N^+(z-x))dF_X(z)dF_X(x)\\
=&\frac{(n-1)(N-n)}{n}\left\{p+\delta_2F_Y'\left(Q(p)\right)+\delta_2^2\frac{F_Y''\left(Q(p)\right)}{2}+o\left(\frac{n}{N^2}\right)\mathbb{I}\left(\delta_2\neq0\right)\right\}\\
&\times\left\{\int\int\left[F_Y\left(Q(p)-\beta_N^+(z-x)\right)+\delta_1F_Y'\left(Q(p)-\beta_N^+(z-x)\right)\right.\right.\\
&\left.\left.+\delta_1^2\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2}+o\left(\frac{n}{N^2}\right)\mathbb{I}\left(\delta_1\neq0\right)\right]dF_X(z)dF_X(x)\right\}\\
=&\frac{(n-1)(N-n)}{n}\left\{p\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right.\\
&+p\times\\
&\int\int\delta_1F_Y'\left(Q(p)-\beta_N^+(z-x)\right)+\delta_1^2\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2}dF_X(z)dF_X(x)\\
&+F_Y'\left(Q(p)\right)\times\\
&\int\int\delta_2F_Y\left(Q(p)-\beta_N^+(z-x)\right)+\delta_1\delta_2F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\\
&\left.+\delta_2^2\frac{F_Y''\left(Q(p)\right)}{2}\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right\}\\
&+o(1)\left[\mathbb{I}\left(\delta_1\neq0\right)+\mathbb{I}\left(\delta_2\neq0\right)\right].
\end{aligned}
\tag{4.38}
$$

Thus

$$
\begin{aligned}
T_6^*=&\frac{(n-1)(N-n)}{n}\left\{-\delta_Np\int\int F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right.\\
&+\delta_NF_Y'\left(Q(p)\right)\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\\
&-\delta_N^2p\int\int\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2}dF_X(z)dF_X(x)\\
&\left.+\delta_N^2\frac{F_Y''\left(Q(p)\right)}{2}\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\right.
\end{aligned}
$$

$$+\delta_N^2 F_Y'\left(Q(p)\right)\int\int F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x)\Big\}+o(1). \tag{4.39}$$

Note that

$$\begin{aligned} T_4^*+T_6^* &= \frac{2(n-1)(N-n)}{n}\delta_N^2 F_Y'\left(Q(p)\right)\int\int F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(z)dF_X(x) \\ &\quad +o(1) \\ &= \frac{2(n-1)(N-n)}{n}\delta_N^2\left[F_Y'\left(Q(p)\right)\right]^2+o(1) \end{aligned} \tag{4.40}$$

using Lemma 4.1.

Recall that $T_7 = O\left(n^{-1}\right)$ and so will not be expanded.

Term (4.25) is

$$\begin{aligned} T_8(q_1,q_2) &= \frac{(N-n)(n-1)}{n^2} \\ &\quad \times\int\int\int F_Y(q_1-\beta_N^+(z-x))F_Y(q_2-\beta_N^+(z-w))dF_X(w)dF_X(z)dF_X(x) \\ &=\frac{(N-n)(n-1)}{n^2}\int\int\int\left[F_Y\left(Q(p)-\beta_N^+(z-x)\right)+o(1)\mathbb{I}\left(\delta_1\neq 0\right)\right] \\ &\quad \times\left[F_Y\left(Q(p)-\beta_N^+(z-w)\right)+o(1)\mathbb{I}\left(\delta_2\neq 0\right)\right]dF_X(w)dF_X(z)dF_X(x) \\ &=\frac{(N-n)(n-1)}{n^2} \\ &\quad \times\int\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)F_Y\left(Q(p)-\beta_N^+(z-w)\right)dF_X(w)dF_X(z)dF_X(x) \\ &\quad +o(1)\left[\mathbb{I}\left(\delta_1\neq 0\right)+\mathbb{I}\left(\delta_2\neq 0\right)\right]. \end{aligned} \tag{4.41}$$

Thus $T_8^* = o(1)$.

Term (4.26) is

$$\begin{aligned} T_9(q_1,q_2) &= \frac{(N-n)(N-n-1)}{n^2} \\ &\quad \int\int\int F_Y\left(q_1-\beta_N^+(z-x)\wedge q_2-\beta_N^+(w-x)\right)dF_X(w)dF_X(z)dF_X(x) \\ &=\frac{(N-n)(N-n-1)}{n^2}\int\int\int\Big[F_Y\left(q_1-\beta_N^+(z-x)\right)\mathbb{I}\left(q_1-q_2\leq\beta_N^+(z-w)\right) \\ &\quad +F_Y\left(q_2-\beta_N^+(w-x)\right)\mathbb{I}\left(q_1-q_2>\beta_N^+(z-w)\right)\Big]dF_X(w)dF_X(z)dF_X(x) \end{aligned}$$

$$
\begin{aligned}
=&\frac{(N-n)(N-n-1)}{n^2} \\
&\times \int\int\int \left\{F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(q_1-q_2\le\beta_N^+(z-w)\right)\right. \\
&\left.+F_Y\left(Q(p)-\beta_N^+(z-w)\right)\mathbb{I}\left(q_1-q_2>\beta_N^+(z-w)\right)\right\} dF_X(w)dF_X(z)dF_X(x) \\
&+o(1)\left[\mathbb{I}\left(\delta_1\neq 0\right)+\mathbb{I}\left(\delta_2\neq 0\right)\right]. \qquad (4.42)
\end{aligned}
$$

Thus

$$
\begin{aligned}
T_9^{**}=&\frac{(N-n)(N-n-1)}{n^2}\left\{\int\int\int\left\{F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(0\le\beta_N^+(z-w)\right)\right.\right. \\
&\left.F_Y\left(Q(p)-\beta_N^+(z-w)\right)+\mathbb{I}\left(0>\beta_N^+(z-w)\right)\right\} dF_X(w)dF_X(z)dF_X(x) \\
&-2\int\int\int\left\{F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(\delta_N\le\beta_N^+(z-w)\right)\right. \\
&\left.+F_Y\left(Q(p)-\beta_N^+(z-w)\right)\mathbb{I}\left(\delta_N>\beta_N^+(z-w)\right)\right\} dF_X(w)dF_X(z)dF_X(x) \\
&+\left\{\int\int\int\left\{F_Y\left(Q(p)-\beta_N^+(z-x)\right)\mathbb{I}\left(0\le\beta_N^+(z-w)\right)\right.\right. \\
&\left.+F_Y\left(Q(p)-\beta_N^+(z-w)\right)\mathbb{I}\left(0>\beta_N^+(z-w)\right)\right\} dF_X(w)dF_X(z)dF_X(x) \\
&+o(1) \\
=&\frac{2(N-n)(N-n-1)}{n^2}\times \\
&\int\int\int\left[F_Y\left(Q(p)-\beta_N^+(z-x)\right)\left(\mathbb{I}\left(0\le\beta_N^+(z-w)\right)-\mathbb{I}\left(\delta_N\le\beta_N^+(z-w)\right)\right)\right. \\
&\left.+F_Y\left(Q(p)-\beta_N^+(z-w)\right)\left(\mathbb{I}\left(0>\beta_N^+(z-w)\right)-\mathbb{I}\left(\delta_N>\beta_N^+(z-w)\right)\right)\right] \\
&\times dF_X(w)dF_X(z)dF_X(x)+o(1) \\
=&o(1) \qquad (4.43)
\end{aligned}
$$

using Lemma 4.2.

Term (4.27) is

$$
\begin{aligned}
T_{10}(q_1,q_2) \;=\; & \frac{(N-n)(N-n-1)(n-1)}{n^2}\int\int F_Y(q_1-\beta_N^+(z-x))dF_X(x)dF_X(z) \\
&\times\int\int F_Y(q_2-\beta_N^+(z-x))dF_X(x)dF_X(z)
\end{aligned}
$$

$$
\begin{aligned}
= & \frac{(N-n)(N-n-1)(n-1)}{n^2} \int\int \left[F_Y\left(Q(p)-\beta_N^+(z-x)\right)\right. \\
& +\delta_1 F_Y'\left(Q(p)-\beta_N^+(z-x)\right)+\delta_1^2\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2} \\
& \left.+o\left(\frac{n}{N^2}\right)\mathbb{I}\left(\delta_1\neq 0\right)\right] dF_X(x)dF_X(z) \\
& \times\int\int\left[F_Y\left(Q(p)-\beta_N^+(z-x)\right)+\delta_2 F_Y'\left(Q(p)-\beta_N^+(z-x)\right)\right. \\
& \left.+\delta_2^2\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2}+o\left(\frac{n}{N^2}\right)\mathbb{I}\left(\delta_2\neq 0\right)\right] dF_X(x)dF_X(z) \\
= & \frac{(N-n)(N-n-1)(n-1)}{n^2} \\
& \times\left\{\left[\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(x)dF_X(z)\right]^2\right. \\
& +(\delta_1+\delta_2)\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(x)dF_X(z) \\
& \times\int\int F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(x)dF_X(z) \\
& +\left(\delta_1^2+\delta_2^2\right)\int\int F_Y\left(Q(p)-\beta_N^+(z-x)\right)dF_X(x)dF_X(z) \\
& \times\int\int\frac{F_Y''\left(Q(p)-\beta_N^+(z-x)\right)}{2}dF_X(x)dF_X(z) \\
& \left.+\delta_1\delta_2\left[\int\int F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(x)dF_X(z)\right]^2\right\} \\
& +o(1)\left[\mathbb{I}\left(\delta_1\neq 0\right)+\mathbb{I}\left(\delta_2\neq 0\right)\right]. \qquad (4.44)
\end{aligned}
$$

Thus

$$
\begin{aligned}
T_{10}^{*} = & \frac{(N-n)(N-n-1)(n-1)}{n^2}\delta_N^2\left[\int\int F_Y'\left(Q(p)-\beta_N^+(z-x)\right)dF_X(x)dF_X(z)\right]^2 \\
& +o(1) \\
= & \frac{(N-n)(N-n-1)(n-1)}{n^2}\delta_N^2\left[F_Y'\left(Q(p)\right)\right]^2+o(1) \qquad (4.45)
\end{aligned}
$$

by Lemma 4.1. Also, note that

$$
\frac{N^2}{n}\left\{F_Y^2(Q_L^+(p_N))-2pF_Y(Q_L^+(p_N))+p^2\right\}
$$

$$
\begin{aligned}
&= \frac{N^2}{n}\Bigg\{\left[p+\delta_N F_Y'\left(Q(p)\right)+\delta_N^2\frac{F_Y''\left(Q(p)\right)}{2}+o\left(\frac{n}{N^2}\right)\right]^2 \\
&\quad -2p\left[p+\delta_N F_Y'\left(Q(p)\right)+\delta_N^2\frac{F_Y''\left(Q(p)\right)}{2}+o\left(\frac{n}{N^2}\right)\right]+p^2\Bigg\} \\
&= \frac{N^2}{n}\delta_N^2\left[F_Y'\left(Q(p)\right)\right]^2+o(1). \qquad (4.46)
\end{aligned}
$$

Now

$$
\begin{aligned}
\mathbb{B}(Q_L^+(p_N);u)-\mathbb{B}(Q(p);u) &= N^{-1}\Big\{n\left(F_Y(Q_L^+(p_N))-p\right) \\
&\quad +(N-n)\left(F_Y(Q_L^+(p_N))-p+O\left(N^{-1}\right)\right)\Big\} \\
&\quad -F_Y(Q_L^+(p_N))+p \\
&= O\left(N^{-1}\right), \qquad (4.47)
\end{aligned}
$$

and $F_Y(Q_L^+(p_N))-p=O\left(n^{\frac{1}{2}}N^{-1}\right)$. Then (4.28) on page 58 is

$$
\begin{aligned}
&\delta_N^2\Bigg\{F_Y'\left(Q(p)\right)\left[n-1+\frac{2(n-1)(N-n)}{n}+\frac{(N-n)(N-n-1)(n-1)}{n^2}-\frac{N^2}{n}\right] \\
&\quad +\frac{u}{\sqrt{n}}\kappa\left[\frac{2(n-1)(N-n)}{n}+\frac{2(N-n)(N-n-1)(n-1)}{n^2}\right] \\
&\quad +o\left(\frac{N}{\sqrt{n}}\right)\Bigg\}+o(1)+O\left(n^{-\frac{1}{2}}\right) \\
&= O\left(n^{-\frac{1}{2}}\right).
\end{aligned}
$$

Thus $Z_N(Q_L^+(p_N);u)-Z_N(Q(p);u)\to_p 0$ and so, for all k and all $\varepsilon>0$,

$$
\begin{aligned}
&\lim_{N\to\infty}\mathbb{P}\left(V_N(u)<k, Z_N(Q(p);u)\geq k+\varepsilon\right) \\
&\quad =\lim_{N\to\infty}\mathbb{P}\left(V_N(u)<k, Z_N(Q_L^+(p_N);u)\geq k+\varepsilon+Z_N(Q_L^+(p_N);u)-Z_N(Q(p);u)\right) \\
&\quad \leq \lim_{N\to\infty}\Bigg\{\mathbb{P}\left(V_N(u)<k, Z_N(Q_L^+(p_N);u)\geq k+\varepsilon-\frac{\varepsilon}{2}\right) \\
&\qquad \times\mathbb{P}\left(\left|Z_N(Q_L^+(p_N);u)-Z_N(Q(p);u)\right|\leq\frac{\varepsilon}{2}\right) \\
&\qquad +\mathbb{P}\Bigg(V_N(u)<k, Z_N(Q_L^+(p_N);u)\geq k+\varepsilon+Z_N(Q_L^+(p_N);u)-Z_N(Q(p);u)\Bigg| \\
&\qquad\qquad \left|Z_N(Q_L^+(p_N);u)-Z_N(Q(p);u)\right|>\frac{\varepsilon}{2}\Bigg) \\
&\qquad \times\mathbb{P}\left(\left|Z_N(Q_L^+(p_N);u)-Z_N(Q(p);u)\right|>\frac{\varepsilon}{2}\right)\Bigg\}
\end{aligned}
$$

$$= \lim_{N\to\infty} \mathbb{P}\left(V_N(u) < k, Z_N(Q_L^+(p_N);u) \geq k + \frac{\varepsilon}{2}\right)$$
$$=0 \tag{4.48}$$

by (4.16). Similarly,

$$\begin{aligned} & \lim_{N\to\infty} \mathbb{P}\left(V_N(u) \geq k, Z_N(Q(p);u) < k+\varepsilon\right) \\ = & \lim_{N\to\infty} \mathbb{P}\left(V_N(u) \geq k, Z_N(Q_L^+(p_N);u) < k+\varepsilon + Z_N(Q_L^+(p_N);u) - Z_N(Q(p);u)\right) \\ \leq & \lim_{N\to\infty} \mathbb{P}\left(V_N(u) \geq k, Z_N(Q_L^+(p_N);u) < k+\frac{3\varepsilon}{2}\right) \\ = & \ 0. \end{aligned} \tag{4.49}$$

Finally, Chambers and Dunstan (1986) showed that $Z_N(Q(p);u)$ has an asymptotic normal distribution. All the conditions for Lemma 1 of Ghosh (1971) are then satisfied and we may conclude that $V_N(u) - Z_N(Q(p);u) \to_p 0$. Noting that $V_N(u) - Z_N(Q(p);u) = Nn^{-\frac{1}{2}} R_N(u)$ and $Nn^{-\frac{1}{2}} \sim n^{\frac{1}{2}}$, we have the result. ▲

Theorem 4.1 is a weaker result than the conjecture of Chambers and Dunstan (1986). Their conjecture is parallel to the result of this theorem but with the random variable, u_n replacing the constant u. Such a replacement requires stronger conditions than $o_p\left(n^{-\frac{1}{2}}\right)$ for $R_N(u)$. For example, uniform convergence on compact sets would suffice to establish the Chambers and Dunstan conjecture. Specifically, if for all $\varepsilon > 0$ and all $M > 0$,

$$\mathbb{P}\left(\sup_{|u|\leq M} n^{\frac{1}{2}} |R_N(u)| > \varepsilon\right) \to 0 \tag{4.50}$$

as $N \to \infty$, then given $\delta > 0$, we can choose M, N so large that

$$\begin{aligned} \mathbb{P}\left(n^{\frac{1}{2}} |R_N(u_n)| > \varepsilon\right) = & \mathbb{P}\left(n^{\frac{1}{2}} |R_N(u_n)| > \varepsilon, |u_n| \leq M\right) \\ & + \mathbb{P}\left(n^{\frac{1}{2}} |R_N(u_n)| > \varepsilon, |u_n| > M\right) \\ \leq & \mathbb{P}\left(\sup_{|u|\leq M} n^{\frac{1}{2}} |R_N(u)| > \varepsilon\right) + \mathbb{P}\left(|u_n| > M\right) \\ < & \frac{\delta}{2} + \frac{\delta}{2}. \end{aligned}$$

Thus $R_N(u_n) = o_p\left(n^{-\frac{1}{2}}\right)$. Establishing (4.50) appears difficult, however.

Theorem 4.1 can be used to obtain an expression for the approximate variance of $\hat{Q}_N(p;u_n)$:

$$\begin{aligned}\mathrm{Var}\left(\hat{Q}_N(p;u_n)\right) &\approx [F'_Y(Q(p))]^{-2}\mathrm{Var}\left[\hat{F}_N(Q(p);u_n)-p\right] \\ &= [F'_Y(Q(p))]^{-2}\Big\{\mathrm{Var}\left[\hat{F}_N(Q(p);u_n)-F_N(Q(p))\right]+\mathrm{Var}[F_N(Q(p))-p] \\ &+ 2\mathrm{Cov}\left(\hat{F}_N(Q(p);u_n)-F_N(Q(p)),F_N(Q(p))-p\right)\Big\}. \qquad (4.51)\end{aligned}$$

Chambers, Dorfman and Hall (1992) gave an expression for the first term and an expression for the second term is well known. However, the third term may be difficult to obtain.

A further linearization of $\hat{Q}_N(p;u)$ will provide an easier way to approximate the variance of $\hat{Q}$ and will make clearer the role played by estimation of β. Using Theorem 4.1, we have that for any fixed u,

$$\hat{Q}_N(p;u) = Q_L(p_N)+\frac{p-\hat{F}_N(Q(p);0)}{F'_Y(Q(p))}+\frac{\hat{F}_N(Q(p);0)-\hat{F}_N(Q(p);u)}{F'_Y(Q(p))}+o\left(n^{-\frac{1}{2}}\right). \qquad (4.52)$$

The next theorem provides a linearized expression for $\hat{F}_N(Q(p);0)-\hat{F}_N(Q(p);u)$.

Theorem 4.2 *If we have condition* C4 *(given on page 55) and*

C6. *for all* t, $F'_E(t)$ *exists,*

then

$$\hat{F}_N(Q(p);u)=\hat{F}_N(Q(p);0)-\frac{\sqrt{n}}{N}\frac{1-\lambda}{\lambda}u\Gamma+R^*_N(u),$$

where $R^*_N(u)=o_p\left(n^{-\frac{1}{2}}\right)$ *and*

$$\Gamma=\mathrm{Cov}\left\{F'_E(Q(p)-X\beta),X\right\}.$$

Proof: By definition (4.6), we have that

$$\begin{aligned}&\frac{N}{\sqrt{n}}(\hat{F}_N(Q(p);u)-\hat{F}_N(Q(p);0)) \qquad (4.53)\\ &=\frac{N}{\sqrt{n}}\frac{1}{Nn}\sum_{i\notin s}\sum_{j\in s}\left[\mathbb{I}\left(X_i\beta+\frac{u}{\sqrt{n}}(X_i-X_j)+E_j\le Q(p)\right)-\mathbb{I}\left(X_i\beta+E_j\le Q(p)\right)\right]\\ &=\frac{N}{\sqrt{n}}\frac{1}{Nn}\sum_{i\notin s}\sum_{j\in s}\mathcal{I}_{ij},\end{aligned}$$

where

$$\mathcal{I}_{ij} = \mathbb{I}\left(X_i\beta + \frac{u}{\sqrt{n}}(X_i - X_j) + E_j \leq Q(p)\right) - \mathbb{I}\left(X_i\beta + E_j \leq Q(p)\right).$$

Define

$$W_{Nj} = n^{-\frac{3}{2}} \sum_{i \notin s_n} \mathcal{I}_{ij}. \tag{4.54}$$

Then

$$\mathbb{E}\left(\sum_{j \in s_n} W_{Nj} \,\middle|\, \{X_j : j \in s_n\}\right) \tag{4.55}$$

$$= \frac{N-n}{n^{\frac{3}{2}}} \sum_{j \in s_n} \int \left[F_E\left(Q(p) - \frac{u}{\sqrt{n}}(z - x_j) - z\beta\right) - F_E(Q(p) - z\beta)\right] dF_X(z)$$

$$= \left(\frac{N-n}{N}\right)\left(\frac{N}{n}\right)\left(\frac{1}{\sqrt{n}}\right) \sum_{j \in s} \int F'_E(Q(p) - z\beta)\left(-\frac{u}{\sqrt{n}}(z - x_j)\right) dF_X(z)$$

$$+ O_p\left(n^{-\frac{1}{2}}\right). \tag{4.56}$$

Then expression (4.56) converges in probability to

$$\frac{1-\lambda}{\lambda}\left(-u \int F'_E(Q(p) - z\beta)(z - \mathbb{E}(X)) dF_X(z)\right)$$

$$= -\frac{1-\lambda}{\lambda}\, u \operatorname{Cov}(F'_E(Q(p) - X\beta), X) = -\frac{1-\lambda}{\lambda}\, u\, \Gamma \tag{4.57}$$

as $N \to \infty$.

We wish to show that (4.53) and (4.55) converge in probability to the same limit. Using definition (4.54), we have that

$$\limsup_{N\to\infty} n\mathbb{E}\left(W_{Nj}^2\right) = \limsup_{N\to\infty} n^{-2} \sum_{i \notin s_n} \sum_{i' \notin s_n} \mathbb{E}\left(\mathcal{I}_{ij}\right)\left(\mathcal{I}_{i'j}\right)$$

$$= \limsup_{N} \Bigg\{ n^{-2}(N-n)\mathbb{E}\left(\mathcal{I}_{ij}\right)^2 + n^{-2}(N-n)(N-n-1)$$

$$\mathbb{E}\Bigg\{\Bigg[F_X\left(\frac{Q(p) - E_j + \frac{u}{\sqrt{n}}X_j}{\beta + \frac{u}{\sqrt{n}}}\right)\mathbb{I}\left(\beta + \frac{u}{\sqrt{n}} > 0\right)$$

$$- \left(1 - F_X\left(\frac{Q(p) - E_j + \frac{u}{\sqrt{n}}X_j}{\beta + \frac{u}{\sqrt{n}}}\right)\right)\mathbb{I}\left(\beta + \frac{u}{\sqrt{n}} < 0\right)$$

$$- F_X\left(\frac{Q(p) - E_j}{\beta}\right)\mathbb{I}\left(\beta > 0\right) - \left(1 - F_X\left(\frac{Q(p) - E_j}{\beta}\right)\right)\mathbb{I}\left(\beta < 0\right)\Bigg]^2\Bigg\}. \tag{4.58}$$

It can be shown that (4.58) is zero. Therefore,

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{j\in s} W_{Nj} - \sum_{j\in s}\mathbb{E}\left(W_{Nj}\,\middle|\,\{X_j : j\in s\}\right)\right)\\
&= \sum_{j\in s}\mathrm{Var}\left(W_{Nj} - \mathbb{E}\left(W_{Nj}\,\middle|\,\{X_j : j\in s\}\right)\right)\\
&= \sum_{j\in s}\mathbb{E}\left(\mathrm{Var}\left(W_{Nj} - \mathbb{E}\left(W_{Nj}\,\middle|\,\{X_j : j\in s\}\right)\,\middle|\,\{X_j : j\in s\}\right)\right)\\
&\leq \sum_{j\in s}\mathbb{E}\left(\mathbb{E}\left(W_{Nj}^2\,\middle|\,\{X_j : j\in s\}\right)\right)\\
&= \sum_{j\in s}\mathbb{E}\left(W_{Nj}^2\right). \qquad (4.59)
\end{aligned}$$

Since (4.58) goes to zero, this quantity goes to zero as $N\to\infty$. Thus

$$\sum_{j\in s_n} W_{Nj} - \sum_{j\in s_n}\mathbb{E}\left(W_{Nj}\,\middle|\,\{X_j : j\in s\}\right)\to 0 \qquad (4.60)$$

in probability as $N\to\infty$. Since $\sum_{j\in s_n} W_{Nj}$ is (4.53), we have that

$$\begin{aligned}
&\frac{N}{\sqrt{n}}\left(\hat{F}_N(Q(p);u) - \hat{F}_N(Q(p);0)\right) + \frac{1-\lambda}{\lambda}\,u\,\Gamma\\
&\quad\sum_{j\in s_n} W_{Nj} - \sum_{j\in s_n}\mathbb{E}\left(W_{Nj}\,\middle|\,\{X_j : j\in s\}\right) + \sum_{j\in s_n}\mathbb{E}\left(W_{Nj}\,\middle|\,\{X_j : j\in s\}\right) + \frac{1-\lambda}{\lambda}u\Gamma,
\end{aligned}$$

which converges in probability to zero as $N\to\infty$, by (4.60) and (4.57). This gives the result. ▲

Using (4.52) and Theorem 4.2, we have that

$$\hat{Q}_N(p;u) = Q_L(p_N) + \frac{p - \hat{F}_N(Q(p);0)}{F_Y'(Q(p))} + \frac{\frac{\sqrt{n}}{N}\frac{1-\lambda}{\lambda}u\Gamma}{F_Y'(Q(p))} + G_N(u), \qquad (4.61)$$

where $G_N(u) = o_p\left(n^{-\frac{1}{2}}\right)$. This expansion is similar to that of Randles (1982), if the conditions are satisfied. However, Randles' conditions are difficult to verify, so a more direct proof was needed.

4.3.2 Variance expression for $\hat{Q}_N(p;u)$

The linearization of $\hat{Q}_N(p;u)$ obtained in (4.61) can be used to derive a large-sample approximation for the variance of the quantile estimator. We now replace u with the

random variable $u_n = \sqrt{n}(\hat{\beta} - \beta)$, where $\hat{\beta}$ is the ordinary least squares estimator of β, to compute the asymptotic variance.

Result 4.1 *Under the conditions of Theorems 4.1 and 4.2* (C1–C5),

$$\begin{aligned}\mathrm{Var}\left(\hat{Q}_N(p;u_n)\right) \approx [F'_Y(Q(p))]^{-2}\Big\{ & n^{-1}(1-\lambda^2) \\ & \times \left[\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x) - p^2\right] \\ & - N^{-1}(1-\lambda)\left[p - \int F_E^2(Q(p)-x\beta)dF_X(x)\right] \\ & + N^{-1}[p-p^2] + (1-\lambda)^2\Gamma^2\mathrm{Var}(\hat{\beta}) \\ & - 2\lambda(1-\lambda)\,\Gamma\,\Omega\,\mathrm{Var}(\hat{\beta})\Big\}, \end{aligned} \tag{4.62}$$

where

$$\Omega = [\mathrm{Var}(E)]^{-1}[\mathrm{Cov}(X, g(X)) + \lambda^{-1}\mathbb{E}(X)\,\mathbb{E}(g(X))]. \tag{4.63}$$

Proof. Using (4.61), we have

$$\begin{aligned}\mathrm{Var}(\hat{Q}_N(p;u_n)) \approx [F'_Y(Q(p))]^{-2}\Big\{ & \mathrm{Var}(p - \hat{F}_N(Q(p);0)) \\ & + \frac{n}{N^2}\frac{(1-\lambda)^2}{\lambda^2}\Gamma^2\mathrm{Var}(u_n) + 2\frac{\sqrt{n}}{N}\frac{1-\lambda}{\lambda}\,\Gamma\,\mathrm{Cov}(p - \hat{F}_N(Q(p);0), u_n)\Big\}. \end{aligned} \tag{4.64}$$

Then the first term of (4.64) is

$$\begin{aligned}&\mathrm{Var}(\hat{F}_N(Q(p);0)) = \\ &\quad \mathrm{Var}\left[\mathbb{E}\left(\hat{F}_N(Q(p);0) \,\middle|\, \{X_i : i \notin s_n\}\right)\right] + \mathbb{E}\left[\mathrm{Var}\left(\hat{F}_N(Q(p);0) \,\middle|\, \{X_i : i \notin s_n\}\right)\right]. \end{aligned} \tag{4.65}$$

The conditional expectation in (4.65) is given by

$$\begin{aligned}&\mathbb{E}(\hat{F}_N(Q(p);0) \,|\, \{X_i : i \notin s\}) \\ &\quad = N^{-1}\mathbb{E}\left(\sum_{j\in s_n}\mathbb{I}(Y_j \le Q(p)) + n^{-1}\sum_{i\notin s_n}\sum_{j\in s_n}\mathbb{I}(E_j \le Q(p) - X_i\beta)\,\middle|\,\{X_i : i \notin s_n\}\right) \\ &\quad = N^{-1}\left[np + \sum_{i\notin s_n}F_E(Q(p) - X_i\beta)\right]. \end{aligned}$$

Thus the first term of (4.65) is

$$\begin{aligned}\mathrm{Var}\left[\mathbb{E}\left(\hat{F}_N(Q(p);0)\,\middle|\,\{X_i : i \notin s_n\}\right)\right] &= N^{-2}(N-n)\mathrm{Var}\left[F_E(Q(p)-X\beta)\right] \\ &= \frac{N-n}{N^2}\left\{\int F_E^2(Q(p)-x\beta)dF_X(x) - p^2\right\}. \end{aligned} \tag{4.66}$$

The conditional variance in (4.65) is given by

$$\begin{aligned}&\mathrm{Var}\left(\hat{F}_N(Q(p);0)\,\middle|\,\{X_i : i \notin s_n\}\right) \\ &= N^{-2}\mathrm{Var}\left(\sum_{j\in s_n}\mathbb{I}(Y_j \le Q(p)) + n^{-1}\sum_{i\notin s_n}\sum_{j\in s_n}\mathbb{I}(E_j \le Q(p)-X_i\beta)\,\middle|\,\{X_i : i \notin s_n\}\right) \\ &= \frac{n}{N^2}\mathrm{Var}\left(\mathbb{I}(E \le Q(p)-Z\beta) + n^{-1}\sum_{i\notin s_n}\mathbb{I}(E \le Q(p)-X_i\beta)\,\middle|\,\{X_i : i \notin s_n\}\right) \\ &= \frac{n}{N^2}\left\{\mathbb{E}\left[\mathbb{I}(E \le Q(p)-Z\beta) + 2n^{-1}\sum_{i\notin s_n}\mathbb{I}[E \le \{(Q(p)-Z\beta)\wedge(Q(p)-X_i\beta)\}]\right.\right. \\ &\quad \left. + n^{-2}\sum_{i\notin s_n}\sum_{i'\notin s_n}\mathbb{I}[E \le \{(Q(p)-X_i\beta)\wedge(Q(p)-X_{i'}\beta)\}]\,\middle|\,\{X_i : i \notin s_n\}\right] \\ &\quad \left. - \left[p + n^{-1}\sum_{i\notin s_n}F_E(Q(p)-X_i\beta)\right]^2\right\} \\ &= \frac{n}{N^2}\left\{p + 2n^{-1}\sum_{i\notin s_n}\int F_E((Q(p)-z\beta)\wedge(Q(p)-X_i\beta))dF_X(z)\right. \\ &\quad + n^{-2}\sum_{i\notin s_n}\sum_{i'\notin s_n}F_E((Q(p)-X_i\beta)\wedge(Q(p)-X_{i'}\beta)) - F_Y^2(Q(p)) \\ &\quad - 2n^{-1}p\sum_{i\notin s_n}F_E(Q(p)-X_i\beta) \\ &\quad \left. - n^{-2}\sum_{i\notin s_n}\sum_{i'\notin s_n}F_E(Q(p)-X_i\beta)F_E(Q(p)-X_{i'}\beta)\right\}. \end{aligned}$$

Thus the second term of (4.65) is

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Var}\left(\hat{F}_N(Q(p);0)\,\middle|\,\{X_i : i \notin s_n\}\right)\right] \\ &= \frac{n}{N^2}\left\{p + 2n^{-1}(N-n)\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x)\right. \\ &\quad + n^{-2}(N-n)p \\ &\quad \left. + n^{-2}(N-n)(N-n-1)\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x)\right. \end{aligned}$$

$$
\begin{aligned}
&- p^2 - 2n^{-1}(N-n)p^2 - n^{-2}(N-n)\int F_E^2(Q(p)-x\beta)dF_X(x) \\
&- n^{-2}(N-n)(N-n-1)p^2\Big\} \\
=&\frac{n}{N^2}\Big\{\left[1+n^{-2}(N-n)\right]p \\
&+\left[-1-2n^{-1}(N-n)-n^{-2}(N-n)(N-n-1)\right]p^2 \\
&+\left[-n^{-2}(N-n)\right]\int F_E^2(Q(p)-x\beta)dF_X(x) \\
&+\left[2n^{-1}(N-n)+n^{-2}(N-n)(N-n-1)\right] \\
&\times\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x)\Big\}. \qquad (4.67)
\end{aligned}
$$

Summing (4.66) and (4.67), we have that (4.65) is

$$
\begin{aligned}
&N^{-2}\Big\{\left[n+\frac{N-n}{n}\right]p+\left[-(N-n)-n-2(N-n)-\frac{(N-n)(N-n-1)}{n}\right]p^2 \\
&+\left[N-n-\frac{N-n}{n}\right]\int F_E^2(Q(p)-x\beta)dF_X(x) \\
&+\left[2(N-n)+\frac{(N-n)(N-n-1)}{n}\right]\times \\
&\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x)\Big\} \\
=&n^{-1}\Big\{\left[1-\frac{n^2}{N^2}-\frac{N-n}{N^2}\right]\times \\
&\left[\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x)-p^2\right] \\
&-\frac{(n-1)(N-n)}{N^2}\left[p-\int F_E^2(Q(p)-x\beta)dF_X(x)\right]\Big\}+\frac{p-p^2}{N}
\end{aligned}
$$

$$
\begin{aligned}
\sim&n^{-1}(1-\lambda^2)\left[\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x)-p^2\right] \\
&-N^{-1}(1-\lambda)\left[p-\int F_E^2(Q(p)-x\beta)dF_X(x)\right] \\
&+N^{-1}[p-p^2] \qquad (4.68)
\end{aligned}
$$

as $N\to\infty$.

The third term of (4.64) involves $\mathrm{Cov}(p - \hat{F}_N(Q(p);0), u_n)$. Since $\hat{\beta}$ is unbiased for β, this covariance is

$$\begin{aligned}
&\mathbb{E}\left(-\sqrt{n}(\hat{\beta}-\beta)N^{-1}\left[\sum_{j\in s_n}\mathbb{I}\left(y_j \le Q(p)\right) + \sum_{i\notin s_n} n^{-1}\sum_{j\in s_n}\mathbb{I}\left(\hat{y}_{ij} \le Q(p)\right)\right]\right) \\
&= -\frac{\sqrt{n}}{N}\mathbb{E}\Bigg\{\mathbb{E}\Bigg[\left(\frac{\sum_{j\in s_n} x_j(x_j+e_j)}{\sum_{j\in s_n} x_j^2} - \beta\right) \\
&\quad\times\left(\sum_{j\in s_n}\mathbb{I}\left(e_j \le Q(p) - x_j\beta\right) + n^{-1}\sum_{i\notin s_n}\sum_{j\in s_n}\mathbb{I}\left(e_j \le Q(p) - x_i\beta\right)\right)\Bigg| X_i : i\in U_N\Bigg]\Bigg\} \\
&= -\frac{\sqrt{n}}{N}\mathbb{E}\Bigg\{\left(\sum_{j\in s_n} x_j^2\right)^{-1}\mathbb{E}\Bigg[\sum_{j\in s_n} x_j e_j \sum_{j'\in s_n}\mathbb{I}\left(e_{j'} \le Q(p) - x_{j'}\beta\right) \\
&\quad + n^{-1}\sum_{j\in s_n} x_j e_j \sum_{i\notin s_n}\sum_{j'\in s_n}\mathbb{I}\left(e_{j'} \le Q(p) - x_i\beta\right)\Bigg| X_i : i\in U_N\Bigg]\Bigg\} \\
&= -\frac{\sqrt{n}}{N}\mathbb{E}\Bigg\{\left(\sum_{j\in s_n} x_j^2\right)^{-1}\left[\sum_{j\in s_n} x_j g(x_j) + n^{-1}\sum_{i\notin s_n}\sum_{j\in s_n} x_j g(x_i)\right]\Bigg\}, \qquad (4.69)
\end{aligned}$$

where

$$g(x) = \int e\,\mathbb{I}\left(e \le Q(p) - x\beta\right) dF_E(e).$$

Then (4.69) is asymptotically

$$\begin{aligned}
&-\frac{\sqrt{n}}{N}[\mathbb{E}(X^2)]^{-1}[\mathrm{Cov}(X, g(X)) + \lambda^{-1}\mathbb{E}(X)\,\mathbb{E}(g(X))] \\
&\sim -\sqrt{n}\,\lambda\Omega\mathrm{Var}(\hat{\beta}), \qquad (4.70)
\end{aligned}$$

where

$$\Omega = [\mathrm{Var}(E)]^{-1}[\mathrm{Cov}(X, g(X)) + \lambda^{-1}\mathbb{E}(X)\,\mathbb{E}(g(X))].$$

Combining (4.68) and (4.70), we have that (4.64) converges to

$$\begin{aligned}
&\left[F_Y'\left(Q(p)\right)\right]^{-2}\Bigg\{n^{-1}(1-\lambda^2) \\
&\qquad\times\left[\int\int F_E((Q(p)-z\beta)\wedge(Q(p)-x\beta))dF_X(z)dF_X(x) - p^2\right] \\
&\qquad - N^{-1}(1-\lambda)\left[p - \int F_E^2(Q(p)-x\beta)dF_X(x)\right] \\
&\qquad + N^{-1}[p-p^2] + (1-\lambda)^2\Gamma^2\mathrm{Var}(\hat{\beta}) - 2\lambda(1-\lambda)\,\Gamma\,\Omega\,\mathrm{Var}(\hat{\beta}).
\end{aligned}$$

Thus we have the result. ▲

Two limiting cases of interest are when the sample size equals the finite population size and when the auxiliary variable is known to be uncorrelated with Y. The first case occurs when $\lambda = 1$. Then Result 4.1 yields

$$\mathrm{Var}(\hat{Q}_N(t; u_n)) \approx \left[F'_Y\left(Q(p)\right)\right]^{-2} N^{-1}\left[p - p^2\right] \tag{4.71}$$

which is simply the model variance of the finite population quantile.

If X and Y are known to be uncorrelated then $\beta = 0$, $\mathrm{Var}(\hat{\beta}) = 0$, and no extra information is provided by the larger sample of X values. Then, using the fact that $n^{-1}\lambda^2 \sim N^{-1}\lambda$, Result 4.1 yields

$$\mathrm{Var}(\hat{Q}_N(t; u_n)) \approx \left[F'_Y\left(Q(p)\right)\right]^{-2} n^{-1}\left[p - p^2\right], \tag{4.72}$$

the usual model variance of the sample quantile.

4.4 Simulation

A small simulation study was designed to evaluate the small sample performance of the asymptotic variance expression given in (4.62). In the study, X is uniformly distributed on $\left(-\frac{1}{2}, \frac{1}{2}\right)$. The distribution of E is $\mathbb{N}(0, \sigma_E^2)$ where σ_E^2 is either 1.0 or 0.02. Two sampling fractions, λ, were used with different sample sizes. Cases I and II have $\lambda = 0.05$. Cases III and IV have $\lambda = .020$. The letters A and B after the case number are used to denote the signal-to-noise ratio $\sigma_E^{-2}\beta^2 = 1.0$ or 5.0. Table 4.1 summarizes the values used in these eight simulation cases.

The asymptotic variance is compared to the empirical variance for the quantile estimator from 100 simulated data sets for each case. The two limiting cases given in (4.71) and (4.72) are also calculated for each case. These can be used as points of reference.

Figures 4.1 and 4.2 show the asymptotic standard deviation, the empirical standard deviation and reference curves for each simulation case. In general, the asymptotic expression performs very well, even for $n = 25$. The empirical standard deviation looks

Table 4.1 Parameter values for simulation.

Case	N	n	λ	β	σ_E^2
IA	500	25	0.05	1.0	1.00
IB	500	25	0.05	1.0	0.20
IIA	1000	50	0.05	1.0	1.00
IIB	1000	50	0.05	1.0	0.20
IIIA	500	100	0.20	1.0	1.00
IIIB	500	100	0.20	1.0	0.20
IVA	1000	200	0.20	1.0	1.00
IVB	1000	200	0.20	1.0	0.20

slightly asymmetric across quantiles, but this is due to using only 100 replications. Using more replications is prohibitive because of computing time.

In all cases, the asymptotic variance lies between the two reference curves. This appears to show that $\hat{Q}$ performs better than a nonparametric quantile estimator from a sample of size n (Q_n), but not as well as that from a sample of size N (Q_N). The asymptotic variance decreases with increased sample size or increased signal-to-noise ratio. The ratio of the asymptotic variance of $\hat{Q}$ to that of Q_n is roughly 70% for all cases in the simulation study, indicating that none of these factors have a significant effect on the efficiency of $\hat{Q}$ relative to Q_n.

4.5 Variance Estimation

4.5.1 Plug-in estimation

Chambers, Dorfman and Hall (1992) gave an expression for

$$\text{Var}\left[\hat{F}_N(q; u_n) - F_N(Q(p))\right]. \tag{4.73}$$

However, as Wu (1999) points out, this expression depends on the model and must be rederived for each new model. Wang and Dorfman (1996) suggest an estimator of (4.73). Wu (1999) suggests a more stable estimate for one term of (4.73). These ideas can be

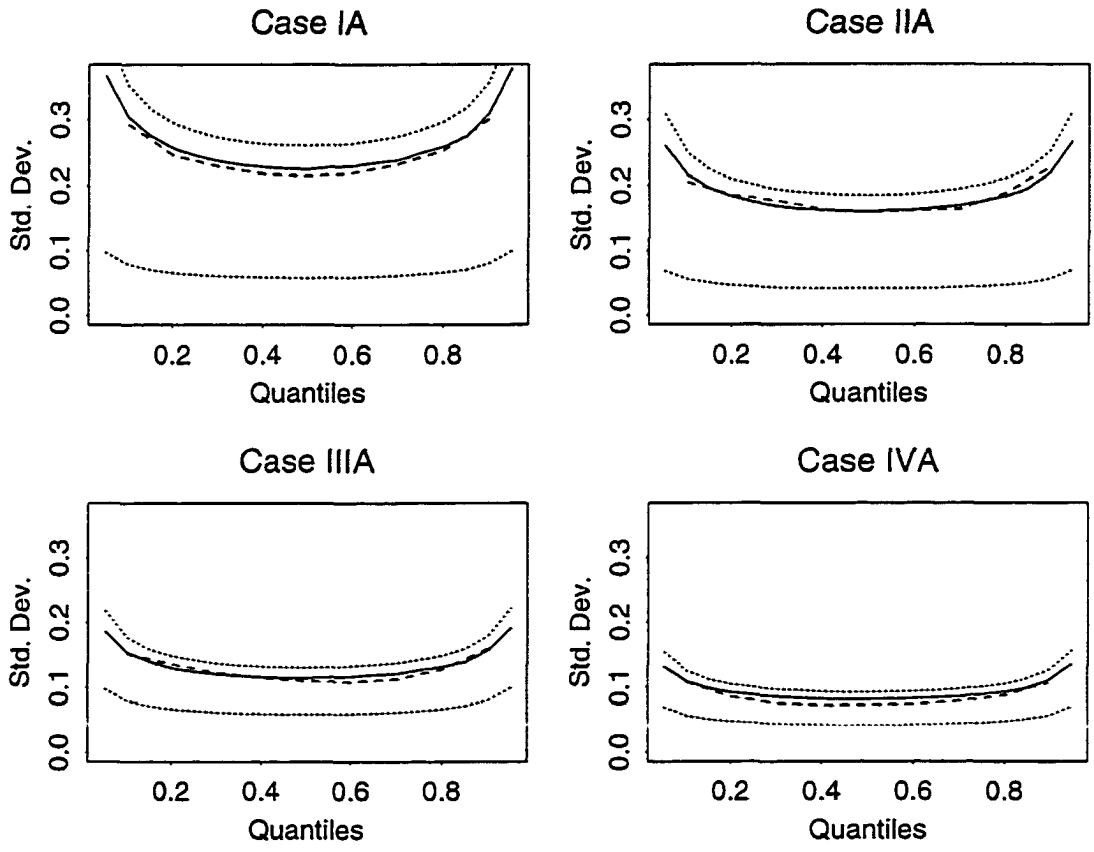

Figure 4.1 Asymptotic and empirical standard deviation for quantile estimators in 'A' cases, with asymptotic standard deviation for empirical quantiles in samples of size n and N for reference. Solid line is asymptotic variance; dashed line is empirical variance from 100 replications; dotted lines are reference curves.

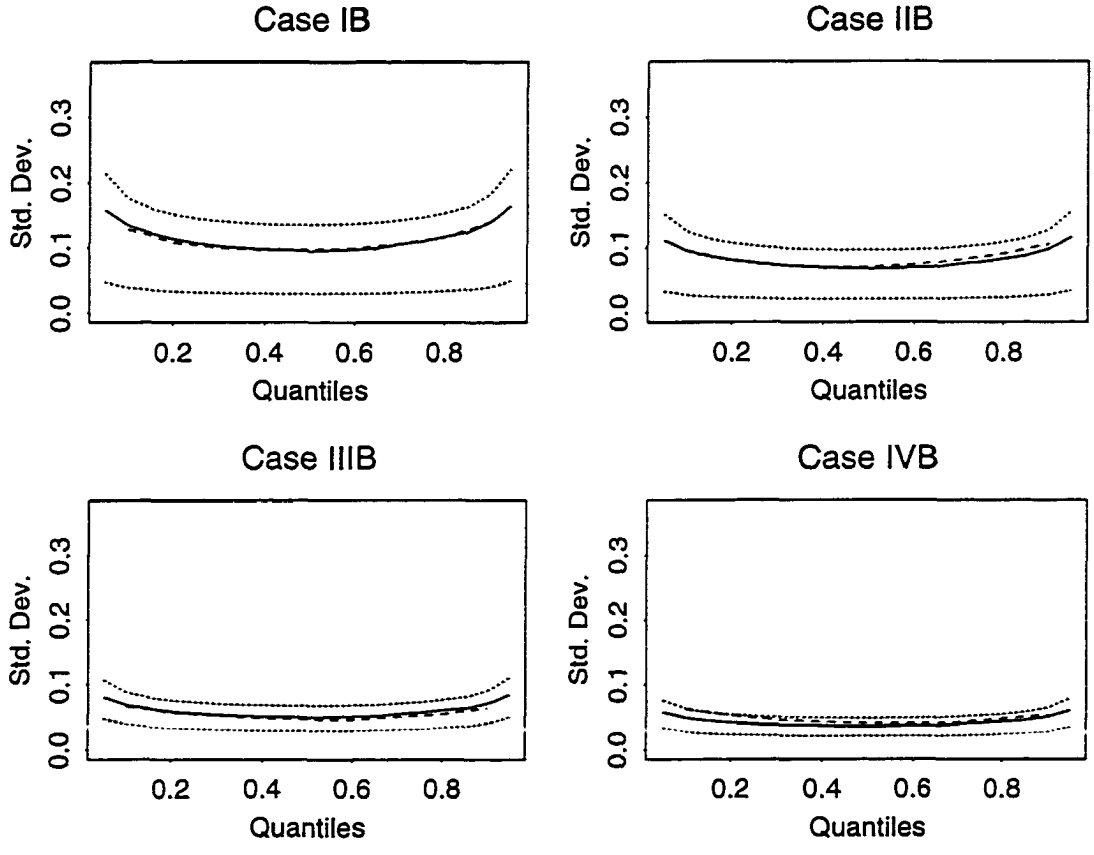

Figure 4.2 Asymptotic and empirical standard deviation for quantile estimators in 'B' cases, with asymptotic standard deviation for empirical quantiles in samples of size n and N for reference. Solid line is asymptotic variance; dashed line is empirical variance from 100 replications; dotted lines are reference curves.

extended to construct a variance estimator for (4.62), v, where

$$
\begin{aligned}
v(\hat{Q}_N(p;u_n)) &= [\hat{f}_Y(Q(p))]^{-2}\Big\{\frac{1-\left(\frac{n}{N}\right)^2}{(N-n)^2} \\
&\times \sum_{i\notin s_n}\sum_{j\notin s_n}\hat{F}_E\big((Q(p)-\hat{\beta}x_i)\wedge(Q(p)-\hat{\beta}x_j)\big)\Big(1-\hat{F}_E\big((Q(p)-\hat{\beta}x_i)\vee(Q(p)-\hat{\beta}x_j)\big)\Big) \\
&- \frac{1-\frac{n}{N}}{N(N-n)^2}\Big[\sum_{i\notin s_n}\hat{F}_E(Q(p)-\hat{\beta}x_i)\Big(1-\hat{F}_E(Q(p)-\hat{\beta}x_i)\Big)\Big] + N^{-1}[p-p^2] \\
&+ \frac{(1-\frac{n}{N})^2 s_E^2}{(N-n)^2 s_X^2}\Big[\sum_{i\notin s_n}(x_i-\bar{x})\hat{f}_E(Q(p)-\hat{\beta}x_i)\Big]^2 \\
&- 2\frac{\frac{n}{N}(1-\frac{n}{N})}{(N-n)^2 s_X^2}\sum_{i\notin s_n}(x_i-\bar{x})\hat{f}_E(Q(p)-\hat{\beta}x_i) \\
&\times\Big[\sum_{i\notin s_n}\big(x_i-(1-\tfrac{n}{N})\bar{x}\big)(Q(p)-\hat{\beta}x_i)\hat{F}_E(Q(p)-\hat{\beta}x_i)\Big]\Big\}, \qquad (4.74)
\end{aligned}
$$

where $\hat{F}_E$ is the empirical distribution function of the fitted residuals, $\hat{f}_Y$ and $\hat{f}_E$ are kernel density estimators, $\bar{x} = N^{-1}\sum_{i\in U_N} x_i$, s_X^2 and s_E^2 are the usual estimates of the variance of X and E, respectively, $\hat{\beta}$ is the ordinary least squares estimate of β and $a \vee b$ denotes the maximum of a and b. This estimate requires non-parametric density estimation. Choice of kernels and bandwidths for estimation of two different densities is a non-trivial problem. In addition, the height of the estimated density and of the estimated distribution function of E are needed at many values. Due to the complexity of this expression, the performance of v is difficult to predict. Instead, we suggest a jackknife approach to variance estimation.

4.5.2 Jackknife variance estimation

Wu (1999) suggests jackknifing to obtain an estimator of (4.73). Shao and Tu (1995, sec. 2.1-2.2) give conditions under which standard jackknife methodology provides a consistent variance estimator. The quantile estimator, $\hat{Q}$, does not meet these conditions because it is not a smooth function. However, Theorem 4.1 suggests that delete-d jackknife methodology may be appropriate (Shao and Tu, 1995, sec. 2.3).

Delete-d jackknife involves deleting d elements from the sample for each jackknife

replicate. If $d = 1$, this is the standard jackknife. The smoothness conditions on the estimator which guarantee consistency of a delete-d jackknife variance estimator are less stringent than those for a standard jackknife. For example, it is well known that the standard jackknife is inconsistent for estimating the variance of a sample quantile. However, the delete-d jackknife is consistent when d is chosen according to

$$\frac{d}{n} \geq \epsilon_0,$$

for some $\epsilon_0 > 0$ and $(n - d) \to \infty$ (Shao and Tu, 1995, p. 53). The proof relies on a Bahadur representation of the sample quantile. Because $\hat{Q}$ is "smoother" than the empirical quantile function, it is anticipated that the delete-d jackknife variance estimator would be consistent in this case as well. While no proof is provided here, simulations in Chapter 7 demonstrate that this is an area worth investigation.

4.6 Extension to two and three phase sampling

The CDE and $\hat{Q}$ can be extended to the case of two-phase sampling. Replace U_N by s_1, the first phase sample, and s_N by s_2, the second phase sample. Let n_1 and n_2 denote the sizes of s_1 and s_2, respectively. For simplicity, we will assume that β is known. Altering notation slightly, we write the CDE for a two phase sample as

$$\hat{F}_2(q) = n_1^{-1}\left[\sum_{j\in s_2} \mathbb{I}(y_j \leq q) + n_2^{-1} \sum_{i\in s_2\setminus s_1} \sum_{j\in s_2} \mathbb{I}(\beta x_i + e_j \leq q)\right]. \tag{4.75}$$

Note that this sum has $n_2 + (n_1 - n_2)n_2$ terms. If we assume that $n_1^{-1}n_2 = \lambda_1 > 0$, then the number of terms is $O(n_1^2)$.

We now consider how the CDE can be extended to three phases. Assume that, as before, we have observations x_i for $i \in s_1$ and y_j for $j \in s_2$. However, we are interested in the distribution of a variable Z which is related to X and Y. Observations z_k are available for $k \in s_3$, a sample of size n_3 from s_2.

4.6.1 Hierarchical modeling of Z

Assume we have a hierarchical model of the form

$$Z = \alpha Y + R \tag{4.76}$$

$$Y = \beta X + E. \tag{4.77}$$

If α and β are known, we observe $\{r_k\}_{k\in s_3}$ and $\{e_j\}_{j\in s_2}$. A natural extension of (4.75) is

$$\hat{F}_3(q) = n_1^{-1}\Bigg[\sum_{k\in s_3}\mathbb{I}(z_k \le q) + n_3^{-1}\sum_{j\in s_2\setminus s_3}\sum_{k\in s_3}\mathbb{I}(\alpha y_j + r_k \le q) + (n_3 n_2)^{-1}\sum_{i\in s_1\setminus s_2}\sum_{j\in s_2}\sum_{k\in s_3}\mathbb{I}(\alpha(\beta x_i + e_j) + r_k \le q)\Bigg]. \tag{4.78}$$

The number of terms in the sum is

$$n_3 + n_3(n_2 - n_3) + n_3 n_2(n_1 - n_2) = O(n_1^3).$$

Thus, extending the CDE through both phases results in a large number of computations. The complexity of the computation can be reduced while still enjoying the benefits of the CDE. If models (4.76) and (4.77) fit well, all fitted residuals from each fitted model may not be necessary to achieve the desired effect. All residuals are not needed to maintain the approximate model-unbiasedness of the estimator. Consider a generalization of $\hat{F}_3$,

$$\hat{F}_3(q; m_2, m_3) = n_1^{-1}\Bigg[\sum_{k\in s_3}\mathbb{I}(z_k \le q) + m_3^{-1}\sum_{j\in s_2\setminus s_3}\sum_{k\in s_3^{(j)}}\mathbb{I}(\alpha y_j + r_k \le q) + (m_3 m_2)^{-1}\sum_{i\in s_1\setminus s_2}\sum_{j\in s_2^{(i)}}\sum_{k\in s_3^{(i)}}\mathbb{I}(\alpha(\beta x_i + e_j) + r_k \le q)\Bigg], \tag{4.79}$$

where $s_a^{(i)}$ is a random sample from s_a of size m_a. The samples $\{s_a^{(i)}\}$ are selected independently for each $i \in s_1 \setminus s_3$. The number of terms in the sum is

$$n_3 + m_3(n_2 - n_3) + m_3 m_2(n_1 - n_2) = O(n_1)O(m_3)O(m_2).$$

For example, if we choose $m_3 = O(1)$ and $m_2 = O(1)$, this is $O(n_1)$. The values of m_2 and m_3 might be chosen based on the fit of models (4.77) and (4.76).

4.6.2 Direct modeling of Z in both phases

Assume that Z, Y and X follow the model

$$Z = \alpha Y + R, \tag{4.80}$$

$$Z = \beta X + E. \tag{4.81}$$

$$\tag{4.82}$$

Under this model, an appropriate extension of (4.75) is

$$n_1^{-1}\Bigg[\sum_{k\in s_3}\mathbb{I}(z_k \le q) + n_3^{-1}\sum_{j\in s_2\setminus s_3}\sum_{k\in s_3}\mathbb{I}(\alpha y_j + r_k \le q) + n_3^{-1}\sum_{i\in s_1\setminus s_2}\sum_{k\in s_3}\mathbb{I}(\beta x_i + e_k \le q)\Bigg]. \tag{4.83}$$

The number of terms in the sum is

$$n_3 + n_3(n_2 - n_3) + n_3(n_1 - n_2) = O(n_1^2).$$

The number of terms in the sum can be reduced in a fashion similar to that of (4.79).

Under some models, a "full" extension of the CDE results in a very large number of terms in the estimator. A generalized version of a CDE extension can be used to reduce the number of calculations needed. It appears that this can be done in a way that maintains most of the advantage of the CDE. More investigation is needed, but is outside the scope of this dissertation.

5 ESTIMATION OF SOIL TEXTURE QUANTILE PROFILES INCORPORATING AUXILIARY INFORMATION

5.1 Introduction

This chapter describes an approach to estimating soil texture quantile profiles. The data structure and notation are described in Chapter 3. The quantile estimator used is of similar form to that presented in Chapter 4. The estimator is modified to accommodate particular features of the data and sampling design. Section 5.2 briefly outlines the estimation procedure. The general approach is to use imputation methodology to predict missing data values. Sections 5.3 through 5.5 describe each step of the procedure in detail. Some examples of estimates are provided in Section 5.5. Jackknife variance estimtes are discussed in Section 5.6. Model diagnostics are discussed in Section 5.7. This estimation approach is compared with a Bayesian approach in Chapter 7.

5.2 Overview of estimation procedure

Recall from Chapter 3 that we wish to obtain estimates of quantile profiles for laboratory determinations of soil texture and corresponding standard errors. Profiles of laboratory texture to 48 inches are available for sites in $\mathcal{L}$. Because of resource constraints, $\mathcal{L}$ by itself is not large enough to support estimation of the distribution of texture for each soil in the county. By design, we have collected auxiliary data which can improve these estimates. Field data for all sites in $\mathcal{D}$ will be combined with laboratory data to improve estimates of distributional parameters.

The multi-phase structure of $\mathcal{D}$ can be viewed as a missing data problem. There are two types of missing data. First, for sites in $\mathcal{S} \cup \mathcal{F}$, laboratory data are missing for observed horizons. Second, for sites in $\mathcal{S}$, full profiles of field and laboratory data are not observed. Two methods, corresponding to the two types of missing data, will be used to predict laboratory values. Table 5.1 summarizes availability of data and predicted values for the sets $\mathcal{S}$, $\mathcal{F}$ and $\mathcal{L}$. The first will be referred to as calibration, the second as imputation, although both are imputation-based methodologies. The second step is motivated by the CDE presented in Chapter 4. The final estimator represents one possible extension of the CDE to a three-phase sample.

Laboratory values are predicted for observed horizons using a calibration model, in which field measurements are used to predict laboratory measurements. A regression model is fit to the field and laboratory data for sites in $\mathcal{L}$ using each horizon as an observation. This model is then used to predict laboratory measurements for all observed field measurements. Predictions from this model are called calibrated values.

Full profiles of calibrated values can be calculated for all sites in $\mathcal{F} \cup \mathcal{L}$. However, for sites in $\mathcal{S}$, calibrated values can be calculated only for the surface horizons. Full profiles for surface horizon sites are the second type of missing data. Imputation will be used to predict texture profiles to a depth of 48 inches for all sites in $\mathcal{D}$.

Table 5.1 Summary of available values. The symbol × denotes the sets for which different types of data and predictions are available.

			$\mathcal{S}$	$\mathcal{F}$	$\mathcal{L}$
Raw data	Field	Surface Horizon	×	×	×
		Full Profile		×	×
	Laboratory	Surface Horizon			×
		Full Profile			×
Predicted laboratory values	Calibrated	Surface Horizon	×	×	×
		Full Profile		×	×
	Imputed	Surface Horizon	×	×	×
		Full Profile	×	×	×

The CDE makes use of complete auxiliary information. That is, auxiliary informa-

tion is available for each element of the population. We will use the calibrated values for the surface horizons as auxiliary information for the calibrated values at greater depths. Complete auxiliary information in this case would mean that a continuous map of surface soil texture was available for the geographic area of interest. However, this is not available. Thus, we adapt the Chambers and Dunstan procedure to the case where partial information is available.

In the original Chambers and Dunstan procedure, the study variable is regressed on the auxiliary data. This model is called the imputation model. Imputed values are then calculated by adding each of the fitted residuals to each model prediction for a missing value. Weights are assigned to each element and a weighted empirical distribution function is calculated from the complete data set (Equation (4.6), page 50).

In the soil texture data, a modified version of the Chambers and Dunstan procedure is applied to the set of calibrated values. Complete auxiliary information is not available, since surface horizon calibrated values are available for sites in $\mathcal{D}$ but not for the entire population. Because the available auxiliary information is derived from a random sample of the population values, imputed values are calculated for all sites in $\mathcal{D}$, including those where full profiles of calibrated values are available. The imputed profiles to a depth of 48 inches for all sites in $\mathcal{D}$ comprise a complete dataset. Weights are calculated for each imputed value. The weights depend on the sampling design, as well as the imputation procedure.

Because the imputed data set has no missing values, standard analysis techniques which incorporate weights can now be applied to the data. However, treating imputed values as observed data does not account for error associated with the predicted value. This problem is well known, especially in the area of variance estimation, where simple imputation methods may produce data that under-represent the amount of variability in the population. (See, for example, Särndal et al., 1991.) The imputation method of the Chambers and Dunstan estimator attempts to address this concern by imputing multiple values for each missing value. Under the imputation model assumptions, the imputed data will have the same amount of variability as the population values.

The complete data can be used to investigate the distribution of texture profiles. This is a joint distribution because it describes probabilities related to the (three-dimensional) texture vector. If we consider the distribution of one component of texture without regard to the values that the other components take, this is a marginal distribution. An estimate of the marginal distribution function of each component of texture s calculated as a weighted empirical distribution function of the back-transformed complete data. Marginal estimates are obtained by inverting this estimate.

5.3 Calibration

We describe the relationship between transformed field and laboratory texture data using a linear model. The fitted model will be used to predict laboratory values, $\boldsymbol{l}_{gh} = (l_{gh,1}, l_{gh,2})$, for observed field data, $\boldsymbol{f}_{gh} = (f_{gh,1}, f_{gh,2})$. Recall that d_{gh} represents the horizon depth of horizon h of site g. The calibration model can be written as

$$l_{gh,1} = \gamma_{10} + \gamma_{11} f_{gh,1} + \gamma_{12} f_{gh,2} + \gamma_{13} h + \gamma_{14} d_{gh} + \gamma_{15} d_{gh} f_{gh,1} + \gamma_{16} d_{gh} f_{gh,2} + \eta_{gh,1}, \tag{5.1}$$

$$l_{gh,2} = \gamma_{20} + \gamma_{21} f_{gh,1} + \gamma_{22} f_{gh,2} + \gamma_{23} h + \gamma_{24} d_{gh} + \gamma_{25} d_{gh} f_{gh,1} + \gamma_{26} d_{gh} f_{gh,2} + \eta_{gh,2}, \tag{5.2}$$

where $\{\eta_{gh,1}\}$ are independent and identically distributed (iid) with mean zero and variance $\sigma^2_{\eta 1}$ and $\{\eta_{gh,2}\}$ are iid with mean zero and variance $\sigma^2_{\eta 2}$. The two error terms, $\eta_{gh,1}$ and $\eta_{gh,2}$, are assumed to be independent of each other. Note that the coefficient γ_{12} allows us to model non-zero correlation between the first component of $\boldsymbol{l}$ and the second component of $\boldsymbol{f}$. Similarly, γ_{21} allows correlation between the second component of $\boldsymbol{l}$ and the first component of $\boldsymbol{f}$. The calibration model can also be written using vector notation as

$$\boldsymbol{l}_{gh} = \boldsymbol{\gamma}' \boldsymbol{f}^{(+)}_{gh} + \boldsymbol{\eta}_{gh},$$

where

$$\gamma' = \begin{pmatrix} \gamma_{10} & \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \gamma_{15} & \gamma_{16} \\ \gamma_{20} & \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26} \end{pmatrix},$$

$$\boldsymbol{f}_{gh}^{(+)} = (1, f_{gh,1}, f_{gh,2}, h, d_{gh}, d_{gh}f_{gh,1}, d_{gh}f_{gh,2})'$$

$$\boldsymbol{\eta}_{gh} = (\eta_{gh,1}, \eta_{gh,2}).$$

In a typical calibration model, the dependent variable is often the measurement which is expected to be subject to greater error. In the soil texture data, it is expected that the laboratory measurement would have less error than the field measurement. However, the data does not indicate that the amount of variability in the laboratory measurements is less than that of the field measurements. Thus, we do not want to treat the laboratory measurements as fixed relative to the field measurements.

Instead, we consider both measurements as random variables from a joint distribution. If their joint distribution were normal, regressing laboratory data on field data would provide us with the best linear unbiased predictor of a laboratory observation given its field observation. We are not assuming normality of the joint distribution of field and laboratory measurements in this analysis, but the reasoning for regressing $\boldsymbol{l}_{gh}$ on $\boldsymbol{f}_{gh}$ is derived from this type of an argument.

To provide a better fit to the data, calibration groups were developed. Using broad landscape-based classes and exploratory data analysis techniques, three groups were chosen: Upland soils (UP), Missouri River Bottom soils (MO) and other River Bottom soils (RB). Let $\mathcal{D}_j$ and $\mathcal{L}_j$ denote the set of sites in $\mathcal{D}$ and $\mathcal{L}$, respectively, that are in calibration group j for $j = 1, 2, 3$. Further classification within these groups was considered, but no other significant groupings were found. To indicate the calibration group, we add a subscript of j to each regression coefficient, γ, and the variances, $\sigma^2_{\eta 1}$ and $\sigma^2_{\eta 2}$.

Data used to fit the models (5.1) and (5.2) are $\{\boldsymbol{l}_{gh}\}$ and $\{\boldsymbol{f}_{gh}\}$ for $h = 1, \ldots, H_g$, for all $g \in \mathcal{L}$. Sample sizes for each calibration group are contained in Table 5.2. The regression coefficients are estimated for each calibration group by ordinary least squares (OLS) regression. The matrix of the estimated coefficients for calibration group j is

denoted $\hat{\gamma}_j$. The variances are estimated with the mean squared errors from the OLS regressions. Let $\hat{\sigma}_{\eta 1 j}$ and $\hat{\sigma}_{\eta 2 j}$ be the estimated variances for calibration group j.

Table 5.2 contains estimates of the regression coefficients and the error variances. These coefficients are for the transformed data, so interpretation of these values is difficult. However the estimates of γ_{11} and γ_{22} are a rough measure of how well calibrated the field and laboratory measurements are to each other. For both of these coefficients, the estimate in the Upland calibration group are very close to 1.0, which indicates good calibration of the two measurement processes. The other groups do not have estimates as close to 1.0 as the Upland group. This is consistent with the soil scientists' assessment of their ability to estimate texture for upland soils relative to Bottomland soils.

Predictions from models (5.1) and (5.2) are called calibrated values and are denoted $\hat{l}_{gh}$, where

$$\hat{l}_{gh} = \hat{\gamma}_j' f_{gh}^{(+)},$$

for $g \in \mathcal{D}_j$. Predictions are calculated for all observed horizons for sites in $\mathcal{D}$. As a simple assessment of the predictive power of the models, we use

$$R^2_{cal}(j,k) = \frac{\sum_{g \in \mathcal{C}_j} \sum_{h=1}^{H_g} (l_{gh,k} - \bar{l}_{jk})^2 - (l_{gh,k} - \hat{l}_{gh,k})^2}{\sum_{g \in \mathcal{C}_j} \sum_{h=1}^{H_g} (l_{gh,k} - \bar{l}_{jk})^2},$$

where

$$\bar{l}_{jk} = \frac{\sum_{g \in \mathcal{C}_j} \sum_{h=1}^{H_g} l_{gh,k}}{\sum_{g \in \mathcal{C}_j} H_g},$$

for $j = 1,2,3$ and $k = 1,2$. The values of $R^2_{cal}(j,k)$ are contained in Table 5.2. See Section 5.7 for more investigation of the fit of this model.

The set of calibrated values has a two phase structure. Calibrated values are available for surface horizons at all sites in $\mathcal{D}$, but full profiles of calibrated values are only available for sites in $\mathcal{F} \cup \mathcal{L}$. In the next section, an imputation procedure is described which will be used to predict full profiles for all sites in $\mathcal{D}$, creating a complete dataset.

Table 5.2 Estimated coefficients, estimated error variances, R^2 values and sample sizes for the calibration models.

Dependent variable	Covariate	Parameter	Estimates for group (j) UP (1)	MO (2)	RB (3)
$\log(c_1^{(l)}/c_3^{(l)})$	Intercept	γ_{10j}	-0.1104	-0.2855	-0.8564
(Model 5.1)	$f_{gh,1}$	γ_{11j}	0.9199	0.7256	0.6078
	$f_{gh,2}$	γ_{12j}	0.0106	-0.0842	-0.3418
	h	γ_{13j}	0.0229	-0.0010	0.0922
	d_{gh}	γ_{14j}	-0.0011	0.0040	0.0046
	$d_{gh}f_{gh,1}$	γ_{15j}	0.0054	0.0056	-0.0011
	$d_{gh}f_{gh,2}$	γ_{16j}	0.0053	0.0049	0.0090
		$\sigma_{\eta 1j}$	0.0674	0.2194	0.0212
		$R^2_{cal}(j,1)$	0.491	0.742	0.618
$\log(c_2^{(l)}/c_3^{(l)})$	Intercept	γ_{20j}	0.2579	-0.5284	-1.0471
(Model 5.2)	$f_{gh,1}$	γ_{21j}	-0.0377	-0.1819	-0.2482
	$f_{gh,2}$	γ_{22j}	1.0706	0.7239	0.6176
	h	γ_{23j}	-0.1897	-0.1226	0.2612
	d_{gh}	γ_{24j}	0.0224	0.0232	-0.0047
	$d_{gh}f_{gh,1}$	γ_{25j}	-0.0001	0.0013	0.0071
	$d_{gh}f_{gh,2}$	γ_{26j}	-0.0021	0.0028	0.0009
		$\sigma_{\eta 2j}$	0.2157	0.8196	0.3261
		$R^2_{cal}(j,2)$	0.767	0.572	0.540
	Number of sites *		84	44	16
	Number of horizons**		508	282	92

* Numbers in this row represent $|\mathcal{L}_j|$.

** Numbers in this row represent $\sum_{g\in\mathcal{L}_j} H_g$.

5.4 Imputation

The CDE was developed for incorporating auxiliary information which is available for the entire population. The CDE can be extended to the case of a two phase sample drawn with unequal selection probabilities. In a two phase sample, auxiliary information is available for a sample of the population and the variable of interest is observed for a subset of this sample. The imputation procedure can be modified to reflect the fact that auxiliary information is not available for the entire population. The estimator is also modified to allow for non-identical distributions among the residuals of the imputation model as in the local residuals estimator studied by Goyeneche (1999). Sampling weights are incorporated into the weighting scheme to account for unequal selection probabilities. These three modifications to the CDE are described below.

Complete auxiliary information is not available in this case. Instead, the auxiliary variables are observed for a random sample of the population. The original CDE is modified by imputing profiles will be calculated for each site in $\mathcal{D}$, including those for which full profiles of calibrated values are available.

The original CD estimator assumes that all residuals come from the same distribution and uses all fitted residuals to impute n values for each missing value. The imputation step can be modified by incorporating imputation classes developed through exploratory data analysis techniques. Only residuals within the same class are used to calculate imputed values for a site. The modified imputation scheme assumes that the distribution of the residuals is the same within each imputation class, but allows different distributions of residuals between classes. See Goyeneche (1999) for details of the properties of this estimator.

The set of calibrated values can be viewed as a two phase sample. Calibrated values for surface horizons are available for all sites in $\mathcal{D}$. However, full profiles of calibrated values are only available for sites in $\mathcal{F} \cup \mathcal{L}$. The calibrated value at the surface is the auxiliary information and the full profile of calibrated values are the variables to be imputed.

The first step of the CD estimator is to impute full profiles of calibrated values

for each site. Observed full profiles have a horizon-based structure as described in Section 3.4. That is, one calibrated value is available for each horizon. The full profile of the horizon structure is not observed for sites in $\mathcal{S}$. Thus, imputing profiles with the same structure as observed full profiles would require imputing the horizon structure as well. Instead, we interpret each observed profile of calibrated values as a piecewise constant function across inches with jumps at the horizon boundaries. This should be a reasonable interpretation because by definition each horizon should have fairly constant properties.

We expand the set of horizon-based calibrated values to a set of inch-based calibrated values. A subscript is added to the calibrated value, so that $\hat{l}_{ghi} = \hat{l}_{gh}$ when inch i is contained in horizon h of site g. See Table 3.2 for an example. Imputed values will be calculated for each inch of a profile.

We will use Old Alluvium soils to demonstrate this step of the estimation procedure. Old Alluvium soils are a subset of the Missouri River Bottom soils calibration group, $\mathcal{D}_2$. Let $\mathcal{D}_{2*}$ denote the set of sites in $\mathcal{D}_2$ that fall on Old Alluvium soils. The set $\mathcal{D}_{2*}$ is partitioned into three imputation classes through exploratory data analysis. The three classes are Luton soils, Keg and Salix soils and other Old Alluvium soils which will be denoted $\mathcal{D}_{2*}^{(1)}$, $\mathcal{D}_{2*}^{(2)}$ and $\mathcal{D}_{2*}^{(3)}$, respectively. In the following, $\mathcal{D}_{2*}$ can be replaced by $\mathcal{D}$ or any subset of $\mathcal{D}$ to extend this method to other subpopulations of interest.

No imputation is needed for $i = 1$; that is, $\hat{l}_{gh1}$ is available for all $g \in \mathcal{D}_{2*}$. We will use the calibrated value from the first inch to predict missing values at greater depths. For $i = 2, \ldots, 48$, we use the models

$$\hat{l}_{ghi,1} = \beta_{10i}^{(j)} + \beta_{11i}\hat{l}_{gh1,1} + \beta_{12i}\hat{l}_{gh1,2} + e_{ghi,1}, \tag{5.3}$$

$$\hat{l}_{ghi,2} = \beta_{20i}^{(j)} + \beta_{21i}\hat{l}_{gh1,1} + \beta_{22i}\hat{l}_{gh1,2} + e_{ghi,2}, \tag{5.4}$$

where site g is in $\mathcal{D}_{2*}^{(j)}$ and $\{e_{ghi,1}\}$ are iid within $\mathcal{D}_{2*}^{(j)}$ with mean zero and variance $\sigma_{eij,1}^2$ and $\{e_{ghi,2}\}$ are iid within $\mathcal{D}_{2*}^{(j)}$ with mean zero and variance $\sigma_{eij,2}^2$. The two error terms, $e_{ghi,1}$ and $e_{ghi,2}$, are assumed to be independent. This model can be written using vector

notation as

$$\hat{l}_{ghi} = \beta_i' \hat{l}_{gh1}^{(+)} + e_{ghi},$$

where

$$\beta_i' = \begin{pmatrix} \beta_{10i}^{(1)} & \beta_{10i}^{(2)} & \beta_{10i}^{(3)} & \beta_{11i} & \beta_{12i} \\ \beta_{20i}^{(1)} & \beta_{20i}^{(2)} & \beta_{20i}^{(3)} & \beta_{21i} & \beta_{22i} \end{pmatrix},$$

$$\hat{l}_{gh1}^{(+)} = \begin{cases} \left(1, 0, 0, \hat{l}_{gh1,1}, \hat{l}_{gh1,2}\right)' & \text{if site } g \text{ is in imputation class 1,} \\ \left(0, 1, 0, \hat{l}_{gh1,1}, \hat{l}_{gh1,2}\right)' & \text{if site } g \text{ is in imputation class 2,} \\ \left(0, 0, 1, \hat{l}_{gh1,1}, \hat{l}_{gh1,2}\right)' & \text{if site } g \text{ is in imputation class 3} \end{cases}$$

and

$$e_{ghi} = \begin{pmatrix} e_{ghi,1} \\ e_{ghi,2} \end{pmatrix}.$$

Data used to fit models (5.3) and (5.4) for inch i are $\{\hat{l}_{ghi}\}$ and $\{\hat{l}_{gh1}\}$ for all $g \in \mathcal{D}_{2*}$ such that $I_g \geq i$. Let

$$n_i^{(j)} = \sum_{g \in \mathcal{D}_{2*}^{(j)}} \mathbb{I}(I_g > i).$$

Then $n_i^{(j)}$ represents the number of observations which are available in imputation class j at inch i for fitting models (5.3) and (5.4). These sample sizes are contained in Table 5.3.

Estimates of the coefficients of models (5.3) and (5.4) are obtained by separate OLS regressions for each inch. The vector of estimated coefficients for inch i is denoted $\hat{\beta}_i$. A selected set of these estimates are contained in Table 5.3. Predictions from models (5.3) and (5.4) are denoted l_{ghi}^* and residuals are denoted $\hat{e}_{ghi}$. That is,

$$\begin{aligned} l_{ghi}^* &= \hat{\beta}_i' \hat{l}_{gh1}^{(+)} \\ \hat{e}_{ghi} &= \hat{l}_{ghi} - l_{ghi}^*. \end{aligned} \tag{5.5}$$

As with the calibration model, we can use R^2 as a simple assessment of the predictive power of the models, where

$$R_{imp}^2(i,k) = \frac{\sum_{g \in \mathcal{D}_{2*}} (\hat{l}_{kghi} - \bar{\hat{l}}_{ik})^2 + (\hat{l}_{kghi} - l_{kghi}^*)^2}{\sum_{g \in \mathcal{D}_{2*}} (\hat{l}_{kghi} - \bar{\hat{l}}_{ik})^2},$$

where

$$\bar{l}_{ik} = \frac{\sum_{g \in \mathcal{D}_{2*}} \hat{l}_{kghi}}{\sum_{j=1}^{3} n_i^{(j)}}$$

for $i = 2, \ldots, 48$; $j = 1, 2, 3$ and $k = 1, 2$. Table 5.3 contains values of $R^2_{imp}(i, k)$. The decreasing trend of the these values across inches reflects the expected pattern in the predictive power of the models. However, R^2 values of 0.580 and 0.616 at (inch 45) are acceptable for biological data. See Section 5.7 for more detailed model assessment.

At each site in $\mathcal{D}_{2*}^{(j)}$, $n_i^{(j)}$ values are imputed for inch i. The imputed values are defined by

$$\tilde{l}_{ghig'} = l^*_{ghi} + \hat{e}_{g'i}$$

for $g, g' \in \mathcal{D}_{2*}^{(j)}$ and $i = 2, \ldots, 48$. Note that only residuals within the same imputation class are used to create imputed values for a site.

We also wish to incorporate sampling weights in the CDE. Each profile in the original dataset has a sampling weight, w_g. At inch i, The sampling weight for each site is partitioned among its $n_i^{(j)}$ imputed values. For $g \in \mathcal{D}_{2*}^{(j)}$, let

$$r_{gi} = \frac{w_g}{\sum_{g' \in \mathcal{D}_{2*}^{(j)}} w_{g'} \mathbb{I}\left(I_{g'} \geq i\right)}.$$

Each imputed value is assigned a weight. For $g, g' \in \mathcal{D}_{2*}^{(j)}$ and $i = 2, \ldots, 48$, the weight is

$$\tilde{w}_{gg'i} = w_g r_{g'i}. \tag{5.6}$$

The imputed data is back-transformed to the original scale using the function $\mathbb{L}^{-1}$ defined in (3.2) (page 43). The back-transformed imputed textures are denoted by $\tilde{c} = (\tilde{c}_{gg'i,1}, \tilde{c}_{gg'i,2}, \tilde{c}_{gg'i,3})$. The weights in (5.6) are used to construct marginal distribution function estimates.

5.5 Estimates

In the original CDE, a weighted empirical distribution function of the complete data is computed as an estimate of the distribution of the variable of interest. Using

Table 5.3 Estimated coefficients, estimated variances, R^2 values, and sample sizes for the imputation models.

Dependent variable	Covariate	Parameter	Inch				
			5	15	25	35	45
$\hat{l}_{ghi,1}$	Intercepts	$\beta_{10i}^{(1)}$	-0.358	3.994	2.150	7.482	12.09
(Model 5.3)		$\beta_{10i}^{(2)}$	-0.384	3.410	1.261	6.258	10.91
		$\beta_{10i}^{(3)}$	-0.374	3.757	1.781	7.073	11.70
	$\hat{l}_{gh1,1}$	β_{11i}	1.028	0.130	0.404	-0.309	-0.786
	$\hat{l}_{gh1,2}$	β_{12i}	-0.119	1.231	0.605	2.301	3.773
		$\sigma^2_{eghi,1}$	0.0021	0.0682	0.0705	0.1435	0.2399
		$R^2_{imp}(i,1)$	0.988	0.708	0.799	0.708	0.616
$\hat{l}_{ghi,2}$	Intercepts	$\beta_{10i}^{(1)}$	-0.120	0.330	0.240	2.140	3.266
(Model 5.4)		$\beta_{10i}^{(2)}$	-0.131	0.142	-0.054	1.781	3.034
		$\beta_{10i}^{(3)}$	-0.127	0.224	0.061	1.958	3.120
	$\hat{l}_{gh1,1}$	β_{21i}	0.009	-0.158	-0.115	-0.353	-0.426
	$\hat{l}_{gh1,2}$	β_{22i}	0.960	1.09	1.034	1.64	2.001
		$\sigma^2_{eghi,2}$	0.0004	0.0118	0.0129	0.0245	0.0287
		$R^2_{imp}(i,2)$	0.983	0.654	0.744	0.634	0.580
		$n_i^{(1)}$	54	14	14	14	13
		$n_i^{(2)}$	30	8	8	8	8
		$n_i^{(3)}$	102	25	25	25	22
Number of imputed values *			14322	3571	3571	3571	3208

* Numbers in this row represent $\sum_{j=1}^{3} n_i^{(j)} n_1^{(j)}$.

the modified weights given in (5.6), we will use a similar estimator for the marginal distribution of each component of texture. The marginal distribution function estimators can be inverted to obtain marginal quantile estimators.

The modified CDE of the marginal distribution of clay at inch i is given by

$$\hat{F}_{i,1}(q) = \frac{\sum_{g,g' \in \mathcal{D}_{2*}} \tilde{w}_{gg'i} \mathbb{I}\left(\tilde{c}_{gg'i,1} \leq q\right)}{\sum_{g,g' \in \mathcal{D}_{2*}} \tilde{w}_{gg'i}} \tag{5.7}$$

The estimators $\hat{F}_{i,2}$ and $\hat{F}_{i,3}$ are defined similarly for sand and silt, respectively. We can also estimate the distribution of clay profiles for any subset of $\mathcal{D}_{2*}$ by only summing over g in the subset.

Note that (5.7) is a step function. This estimator is transformed to a continuous function by connecting the midpoints of the rises of the step function using a procedure described in Nusser et al (1996). Continuous versions of $\hat{F}_{i,1}$, $\hat{F}_{i,2}$ and $\hat{F}_{i,3}$ are denoted by $\tilde{F}_{i,1}$, $\tilde{F}_{i,2}$ and $\tilde{F}_{i,3}$.

By inverting the marginal distribution function estimators, we obtain marginal quantile estimators, $\tilde{Q}_{i,1}$, $\tilde{Q}_{i,2}$ and $\tilde{Q}_{i,3}$. For example, the estimator of the pth quantile of clay at inch i is

$$\tilde{Q}_{i,1}(p) = \inf\{q : \tilde{F}_{i,1}(q) \geq p\}.$$

The quantile estimators are smoothed across inches to obtain the final quantile estimates, $Q^*_{i,1}$, $Q^*_{i,2}$ and $Q^*_{i,3}$. For example, for $i = 3, ...46$, the smoothed estimators are

$$Q^*_{i,1}(p) = \frac{1}{5} \sum_{i'=i-2}^{i+2} \tilde{Q}_{i',1}(p),$$

and to smooth the estimates at the ends,

$$Q^*_{1,1}(p) = \hat{Q}_{3,1}(p)$$
$$Q^*_{2,1}(p) = \hat{Q}_{3,1}(p)$$
$$Q^*_{47,1}(p) = \hat{Q}_{46,1}(p)$$
$$Q^*_{48,1}(p) = \hat{Q}_{46,1}(p).$$

Figure 5.1 contains quantile profile estimates with data for Old Alluvium soils. Each data point is plotted using the depth at the mid point of the horizon. Across inches,

40% of the data points fall in the estimated interquartile range; 72% of the data points fall between the .10 and .90 quantiles; 83% of the data points fall between the .05 and .95 quantiles; and 94% of the data points fall between the .01 and .99 quantiles. These frequencies are only a rough guideline for assessing the estimated quantiles, but they suggest that the estimated distribution may be too peaked relative to the true distribution of laboratory measurements. In Chapter 7, we will investigate this further.

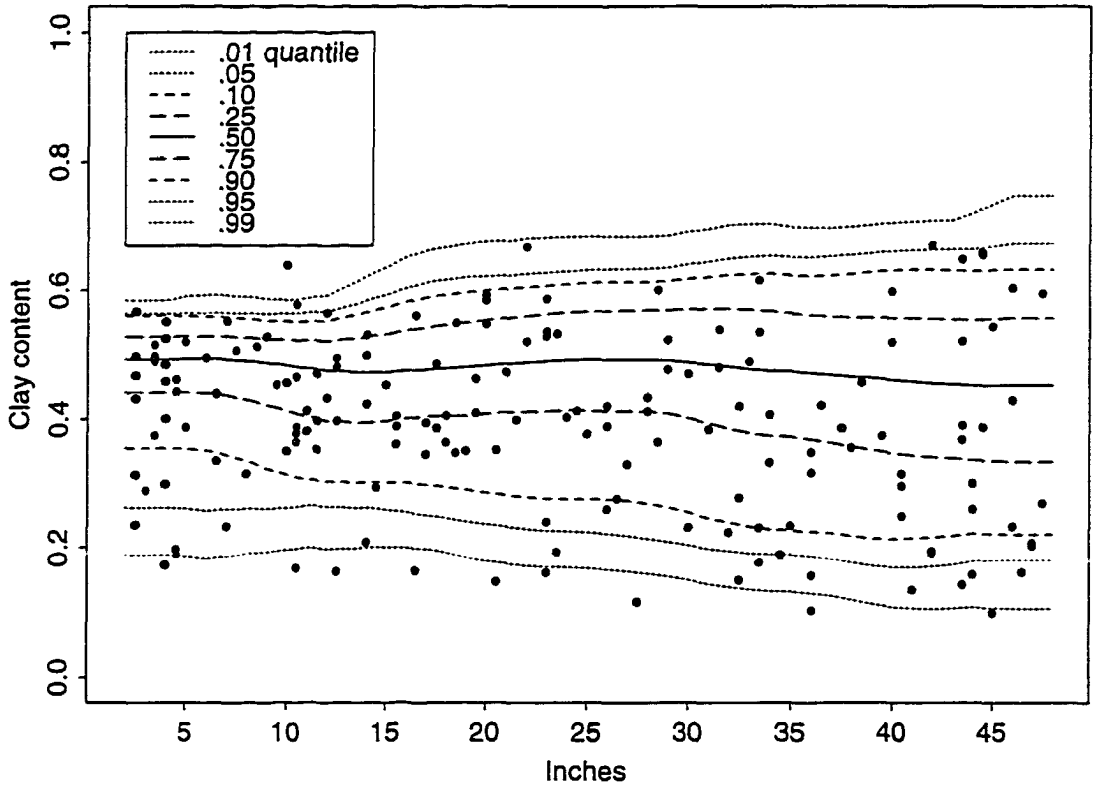

Figure 5.1 Clay quantile estimates for Old Alluvium soils with data. The horizontal axis represents inches; the vertical axis represents clay content. The solid, dashed and dotted lines are quantiles as indicated in the legend. Dots on the graph represent laboratory determinations of clay content for each horizon at each site $g \in \mathcal{L}$. Each dot is plotted at the midpoint of the horizon.

5.6 Variance estimation

A delete-d jackknife procedure is used to estimate the variance of the quantile profiles. The appropriateness of this methodology is suggested by theoretical results from Chapter 4 and simulation results from Chapter 7. The basic idea is to delete d elements from the sample for each jackknife replicate. The purpose of deleting d elements is to overcome the lack of smoothness in the estimator. The less smooth the estimator, the larger d must be. In general, this requires $\begin{pmatrix} n \\ d \end{pmatrix}$ jackknife replicates. However, certain subsets of all possible combinations can be used to approximate the delete-d jackknife.

Implementation of a delete-d jackknife for a multi-phase sample is not necessarily straightforward. We have chosen a procedure which seems intuitively reasonable, but needs further theoretical investigation in the future. We construct clusters of size four because of the sampling rate for the phase 2 sample. A cluster consists of three sites in $\mathcal{S}$ and one site in either $\mathcal{F}$ or $\mathcal{L}$. Clusters are paired to create strata.

For each jackknife replicate, one cluster is deleted from the imputed data set. That is, the calibration and imputation models are fit using the full data set and the coefficients in these models are not recalculated for each jackknife replicate. When the jkth cluster is deleted, we delete all $\tilde{l}_{ghig'}$ such that g or g' are in the jkth cluster. The resulting variance estimator will not reflect the variability in $\hat{Q}$ due to estimation of the coefficients in the calibration and imputation models. Thus it will be negatively biased. However, the computing time is less and it is expected that the bias will not be significant.

The weights for the imputed values generated from sites in the remaining cluster in the jth stratum are reweighted to compensate for the deleted cluster. This results in doubling the weight for $\tilde{l}_{ghig'}$ g or g' is in the remaining cluster and quadrupling the weight for imputed values such that g and g' are in the remaining cluster.

A jackknife variance estimate is obtained by calculating the variability of the reduced estimates relative to the parent estimates. For example, the variance estimate for the

pth quantile of clay for imputation class l at inch i is

$$v_{i,1} = \sum_{j,k} 0.5 \left(\hat{Q}_{(jk),1}(p) - \hat{Q}_{(\cdot k),1}(p) \right)^2,$$

where $\hat{Q}_{(jk),1}(p)$ is the estimate of the pth clay quantile when cluster jk is deleted and

$$\hat{Q}_{(\cdot k),1}(p) = 0.5 \sum_{k} \hat{Q}_{(jk),1}.$$

Figure 5.2 shows jackknife standard deviation profiles for three quantiles. The standard deviation for the lower quartile is much higher than that of the other two quantiles. If the shape of the true distribution is left-skewed, as the estimated quantile profiles suggest, then this pattern is not surprising. The estimated standard deviation is around 0.01 for the median for most of the profile. Using a normal approximation, we can say that we have estimated the median of clay content to within 2% clay.

5.7 Model assessment

The performance of the estimators presented in Section 5.5 is affected by how well the calibration and imputation models fit the data. In Tables 5.2 and 5.3, R^2 values were presented as a simple assessment of the predictive power of the models. In this section, we present a more detailed assessment of the fit of these models.

5.7.1 Calibration models

Scatter plots of the transformed laboratory and field determinations are contained in Figure 5.3. Each component of $\boldsymbol{l}_{gh}$ is plotted against each component of $\boldsymbol{f}_{gh}$ for each calibration group. While some of the plots indicate nonlinear relationships, in general, linear models should capture most of the trends. Fitted residuals from each component of the calibration model and for each calibration group are plotted against each of the covariates in Figures 5.4 and 5.5. These plots do not show any serious problems with the model. Although, we note that the variance of the residuals is not the same for all calibration groups or for both components of the error vector. The calibration model seems to be reasonably satisfied for this data.

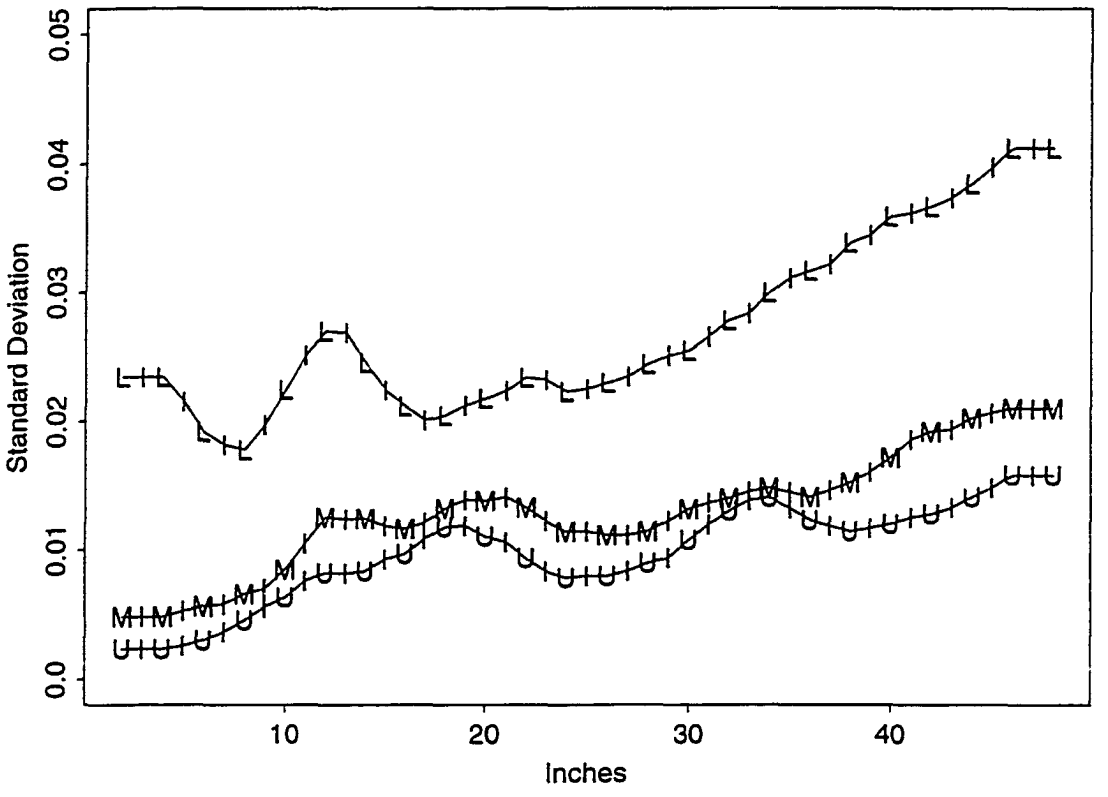

Figure 5.2 Estimated standard deviation profiles for the median ('M'), the lower quartile ('L') and the upper quartile ('U') from the jackknife procedure.

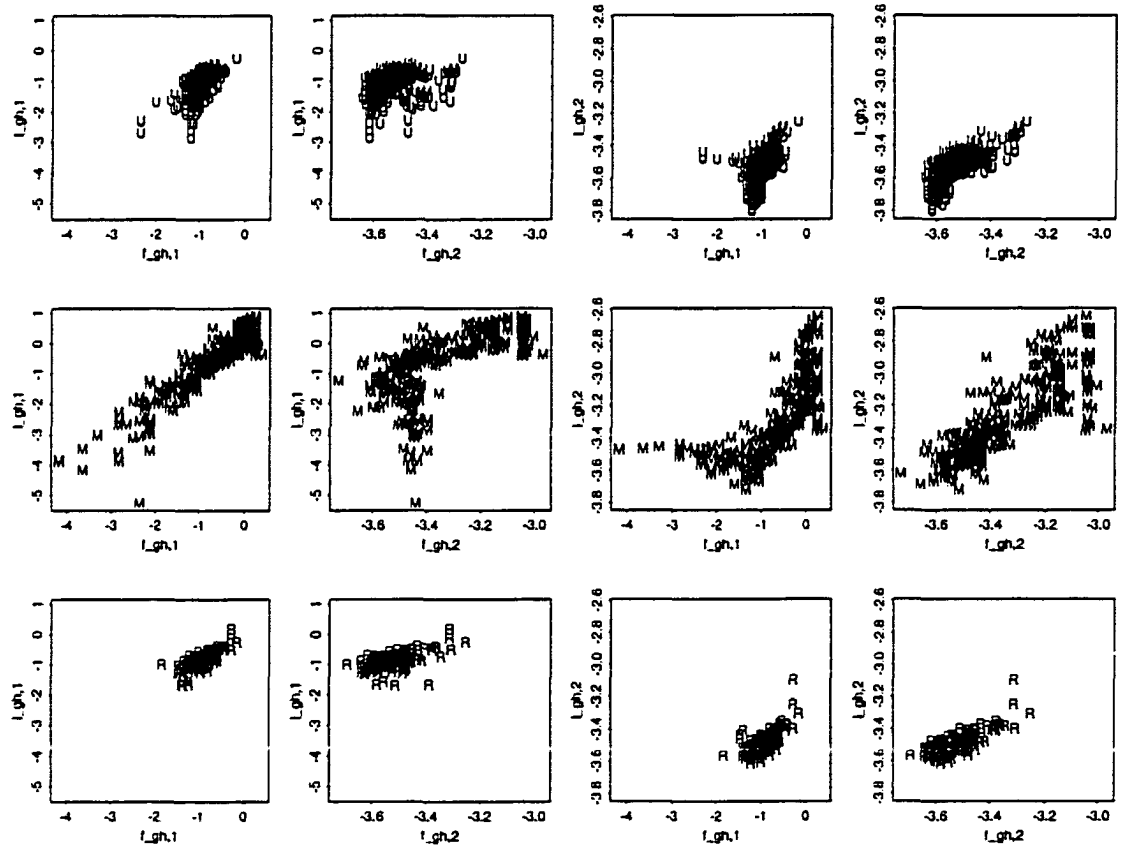

Figure 5.3 Scatter plots for each calibration group: Upland (first row), Missouri River Bottom (second row) and other River Bottom (third row) of $f_{gh,1}$ and $l_{gh,1}$ (first column); $f_{gh,2}$ and $l_{gh,1}$ (second column); $f_{gh,1}$ and $l_{gh,2}$ (third column); and $f_{gh,2}$ and $l_{gh,2}$ (fourth column). Plotting symbols represent calibration groups.

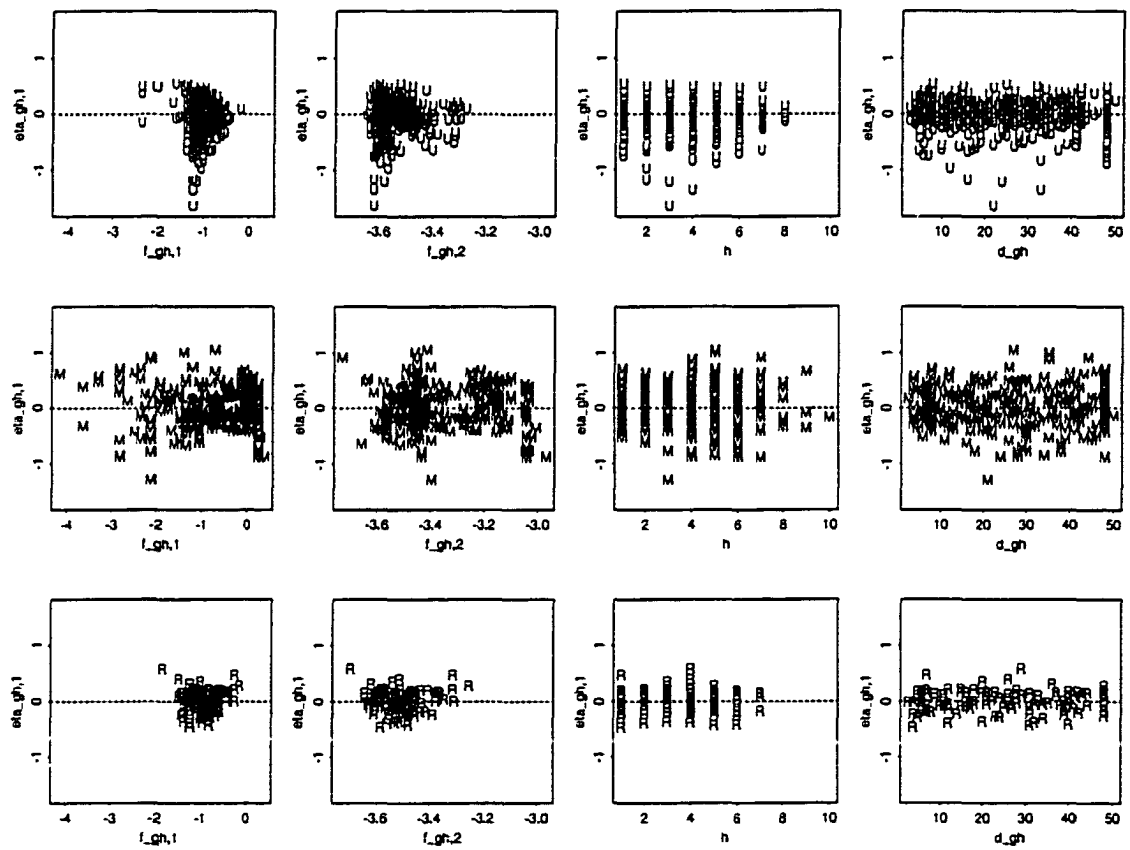

Figure 5.4 Residuals ($l_{gh,1} - \hat{l}_{gh,1}$) for each calibration group: Upland (first row), Missouri River Bottom (second row) and other River Bottom (third row) versus $f_{gh,1}$ (first column); versus $f_{gh,2}$ (second column); horizon sequence, h (third column); and horizon depth, d_{gh} (fourth column). Plotting symbols represent calibration groups.

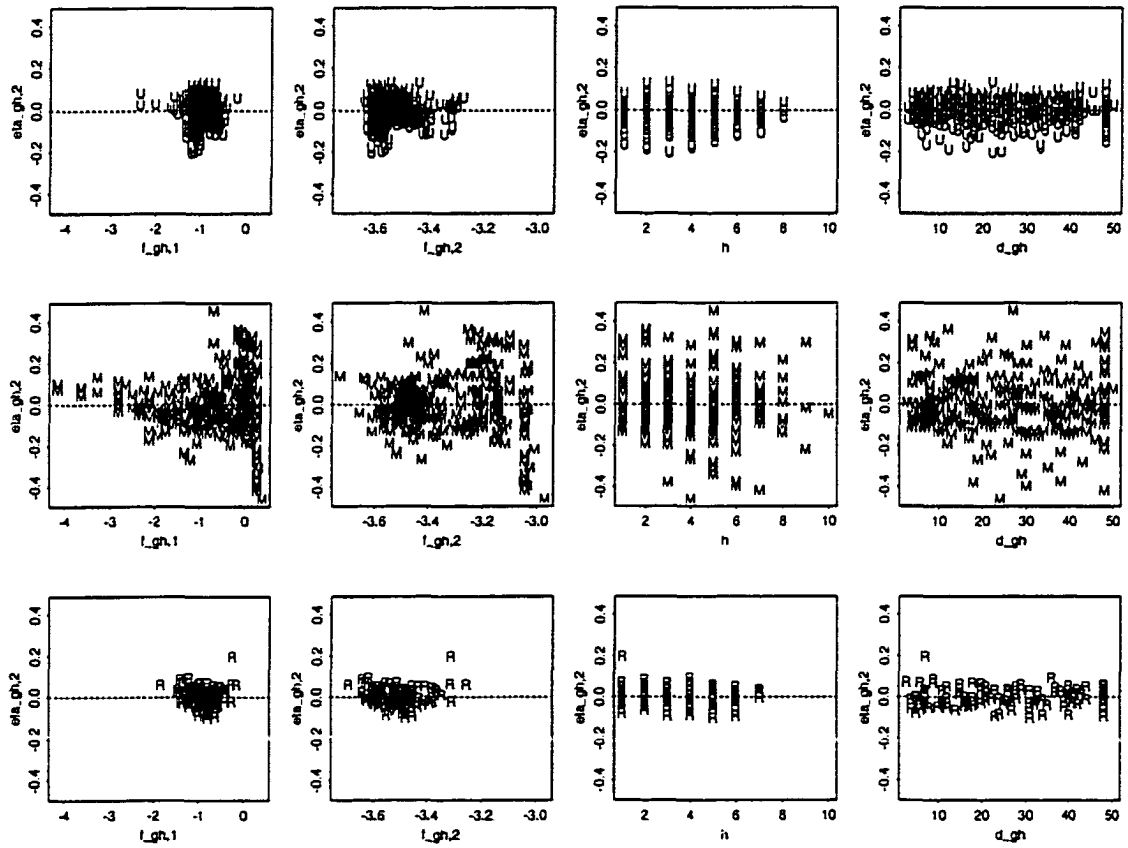

Figure 5.5 Residuals ($l_{gh,2} - \hat{l}_{gh,2}$) for each calibration group: Upland (first row), Missouri River Bottom (second row) and other River Bottom (third row) versus $f_{gh,1}$ (first column); versus $f_{gh,2}$ (second column); horizon sequence, h (third column); and horizon depth, d_{gh} (fourth column). Plotting symbols represent calibration groups.

5.7.2 Imputation models

As with the calibration models, we consider scatter plots of each component of the dependent variable on each component of the covariate. These plots are contained in Figures 5.6 and 5.7 for selected inches for all imputation classes. The trends are fairly linear, although we see a deterioration in the strength of the relationship from the surface to the bottom of the profile (left to right in the figures).

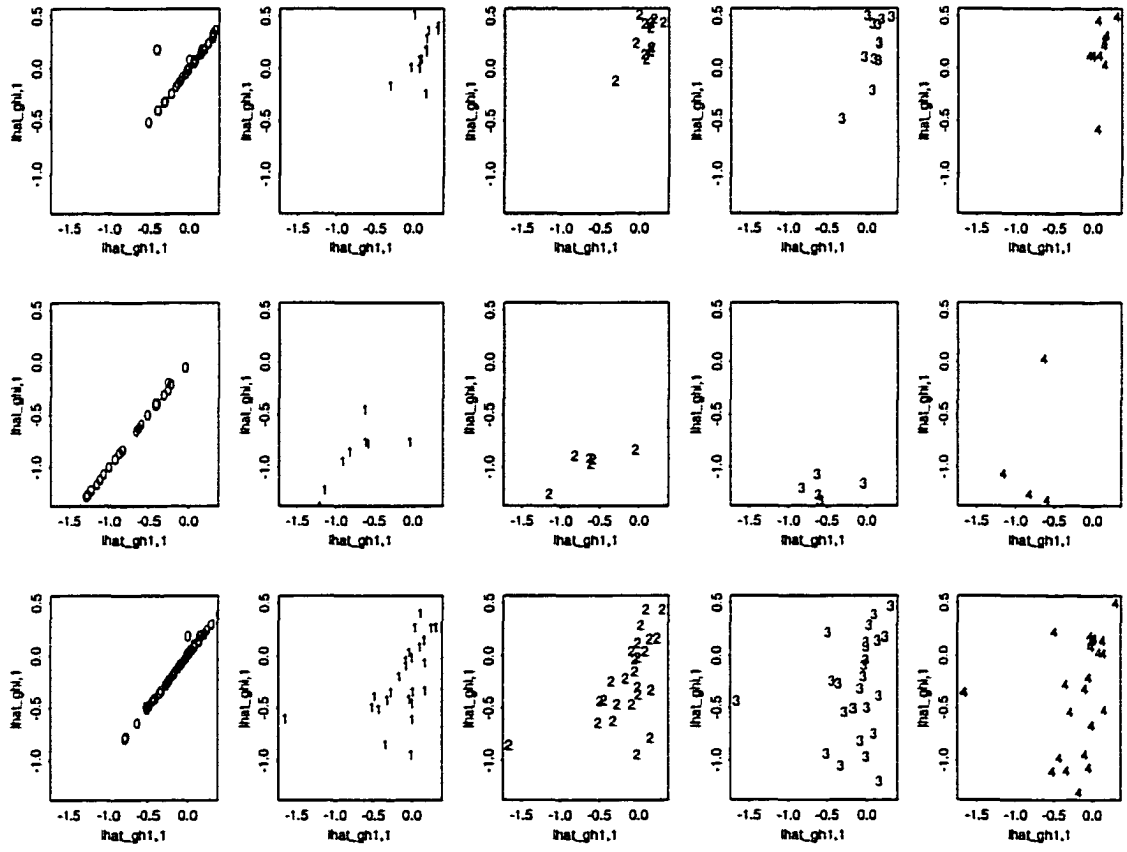

Figure 5.6 For each imputation class: Luton (first row), Keg/Salix (second row) and Other Old Alluvium soils (third row), scatter plot of $\hat{l}_{gh1,1}$ versus $\hat{l}_{ghi,1}$ for $i = 5, 15, 25, 35, 45$ (columns 1, 2, 3, 4, 5, respectively). Plotting symbols represent the integer part of $i/10$.

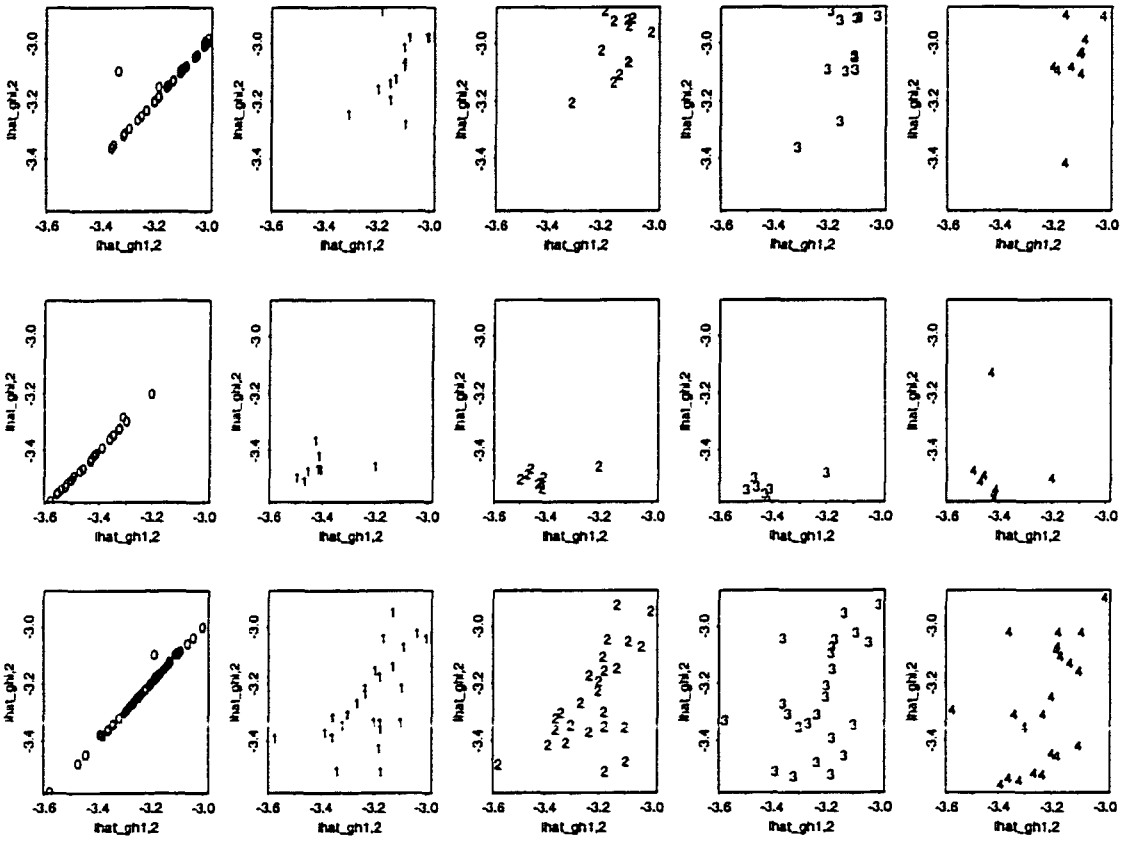

Figure 5.7 For each imputation class: Luton (first row), Keg/Salix (second row) and Other Old Alluvium soils (third row), scatter plot of $\hat{l}_{gh1,2}$ versus $\hat{l}_{ghi,2}$ for $i = 5, 15, 25, 35, 45$ (columns 1, 2, 3, 4, 5, respectively). Plotting symbols represent the integer part of $i/10$.

The imputation models allow a different intercept for each imputation class, but restrict the slope coefficients to be the same across imputation classes. This decision was made because of the instability of the estimated regression coefficients for a model without this restriction. The instability arose from the small sample size below 10 inches in each imputation class. The restriction brings more stability to the coefficient estimates by pooling the imputation classes together in order to estimate the coefficients of the components of $\boldsymbol{f}_{gh}$. A model with different slopes for each imputation class, but the same intercept was also considered. However, the selected model fits slightly better and has more justification from a soil science point of view.

To demonstrate the conclusions from residual plots, one set of plots is included in Figure 5.8. Other plots exhibit similar behavior. In general, residuals in the imputation class other Old Alluvium soils have larger variance than the other two classes. It is difficult to assess the size of the variance in the Keg/Salix class relative to the other two because of the small sample sizes. The difference in variances between the imputation classes suggests that the local residuals imputation is likely to perform better than the original imputation of the CDE. The deterioration of the strength of the linear relationship between surface calibrated values and calibrated values deeper in the profile is reflected in the residual plots. The variance of the residuals increases across inches.

5.8 Conclusion

Estimated quantile profiles for laboratory measurements of soil texture data are desired for the soil texture data collected in the MLRA 107 pilot project. The sample was designed to take advantage of auxiliary information in the form of field measurements. Modifications to the quantile estimator of Chapter 4 are needed to accommodate features of the pilot project data. Assumptions of the calibration and imputation models appear to be reasonably satisfied. However, comparing estimated quantile profiles to observed data indicates that this methodology may produce estimated distributions which are too peaked. This phenomenon is also seen using simulated data in Chapter 7 and a random calibration step is proposed to combat the bias.

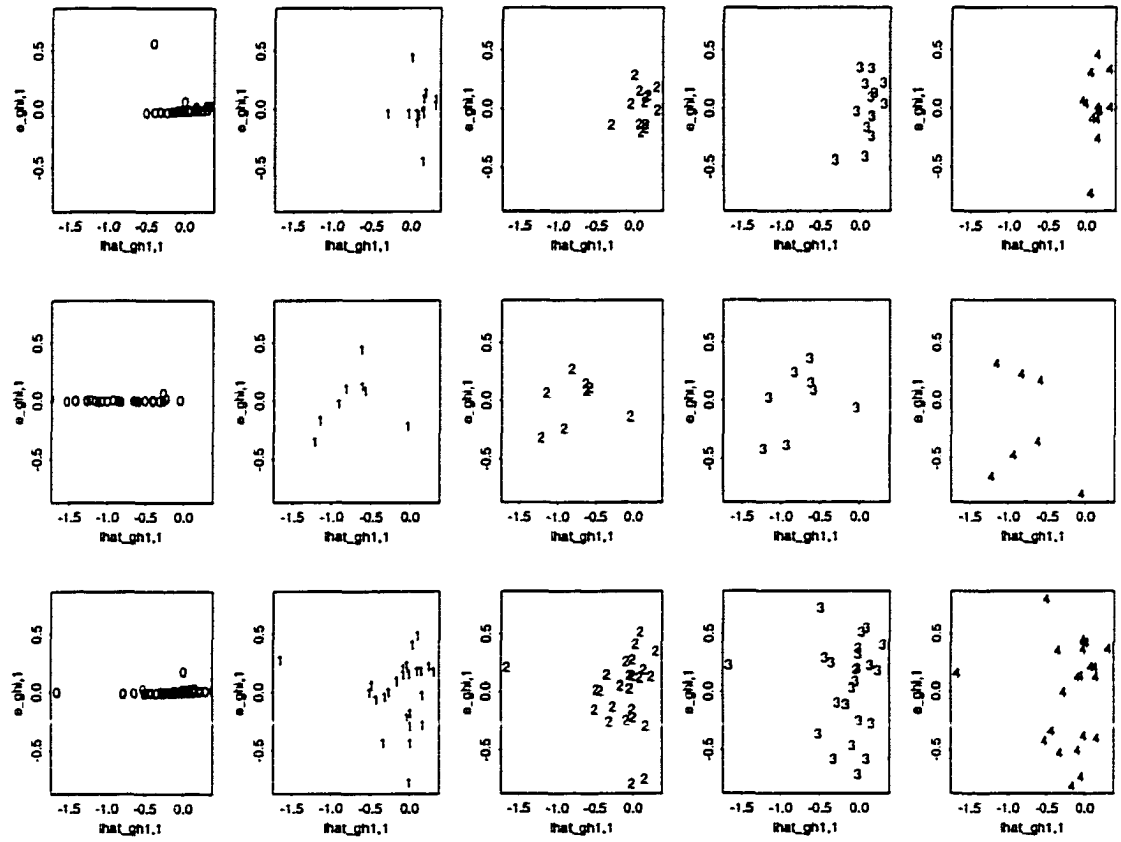

Figure 5.8 For each imputation class: Luton (first row), Keg/Salix (second row) and Other Old Alluvium soils (third row), residuals $(l^*_{ghi,1} - \hat{l}_{ghi,1})$ versus $\hat{l}_{gh1,1}$ for $i = 5, 15, 25, 35, 45$ (columns 1, 2, 3, 4, 5, respectively). Plotting symbols represent the integer part of $i/10$.

6 ESTIMATION OF SOIL TEXTURE QUANTILE PROFILES USING A HIERARCHICAL MODEL

In this chapter, we introduce a hierarchical model to analyze the texture profile data introduced in Chapter 3. Maximum likelihood estimates are derived for the simplest version of the model and Bayesian inference is used in the general case. The Bayesian methodology provides a unified solution to the problems of parameter estimation, imputation of missing data, prediction of new profiles, and so forth, and explicitly accounts for uncertainty through generation of posterior distributions.

Gibbs sampling is used to generate a numerical approximation to the posterior distribution of the parameters given the data. A brief introduction to relevant methods is contained in Section 6.1. See Gelman et al. (1995) for a thorough treatment of Bayesian methodology. Sections 6.2 through 6.6 describe the model and posterior distributions. Posterior profiles are defined in Section 6.3. To assess the adequacy of the model, marginal posterior distributions are compared with estimates from classical analysis in Section 6.8. Other methods of model checking in that section suggest some aspects of the model which may not be adequate. Possible extensions of the model which address these areas are presented in Section 6.9. The results of this analysis are compared to the imputation approach of Chapter 5 in Chapter 7.

6.1 Introduction to Bayesian methodology

6.1.1 Bayesian inference

Bayesian inference involves the use of a probability model which describes the structure of relationships among the data and parameters relevant to the data. The primary

tool for making inference is the posterior distribution which is derived via Bayes' rule. The parameters, θ, are usually unobservable and unknown, so a prior distribution is used to describe the researcher's uncertainty about the value of the parameters. The prior distribution of the parameters is denoted $p(\theta)$. The data model describes the distribution of the observable data, y, given values of the parameters. The data model is denoted $p(y \mid \theta)$. Bayes' rule states that

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{\int_\theta p(\theta)p(y \mid \theta)}.$$

The distribution $p(\theta \mid y)$ is called the posterior distribution.

Both the prior distribution and the data model should reflect what can reasonably be believed about the relationships of the observable and unobservable quantities. Prior distributions may be specified based on some prior information which is available about the parameters. Often, prior information is not available or is difficult to translate into a prior distribution. In this case, we may choose a prior distribution which has support on any believable value of the parameters and which is sufficiently vague. By vague, we mean a distribution which does not place "too much" mass on any particular set of values.

As in classical analysis, the form of the data model can be difficult to choose. Again, prior information may be useful in specifying the form of the model. Model checking methods are available for assessing the plausibility of a proposed model. Assessment of the model presented in this chapter is considered in Section 6.8.

6.1.2 Hierarchical models

The type of probability model used for this analysis is a hierarchical model. Hierarchical models are often used when the parameters of a model are thought to be related in some way. Suppose the data model is specified as $p(y \mid \theta)$. The elements of the parameter vector, θ, are believed to have a dependence structure which we wish to specify in the prior distribution. For a complex multi-parameter problem, it may be convenient to conceptualize this dependence through another set of parameters, ϕ, called hyperpa-

rameters. Then $p(\theta) = p(\theta \mid \phi)p(\phi)$, where $p(\phi)$ is called the hyperprior distribution. The posterior distribution now also includes ϕ. That is,

$$p(\theta, \phi \mid y) = \frac{p(y \mid \theta)p(\theta \mid \phi)p(\phi)}{\int_{\theta,\phi} p(\phi)p(\theta \mid \phi)p(y \mid \theta)}.$$

In general, we will not distinguish between parameters and hyperparameters. Thus the posterior distribution will continue to be denoted $p(\theta \mid y)$, even though some elements of θ might be considered hyperparameters.

6.1.3 MCMC methods

In many settings, the posterior distribution is a multi-dimensional non-standard distribution. This is the case for the model presented in Sections 6.2 and 6.5. This means that, while one may be able to write down the form of the posterior distribution, deriving closed form formulas for summaries of the distribution such as means and variances may be impossible. In these cases, simulation methods may be used to generate pseudo-random draws $\theta^{(t)}$ from the posterior distribution $p(\theta \mid y)$. The posterior distribution can then be summarized using simple numerical techniques. For example, posterior means of parametric functions $h(\theta)$ are approximated as $M^{-1} \sum_{k=1}^{M} h(\theta^{(t)})$.

A commonly used class of simulation methods is called Markov chain Monte Carlo (MCMC). See, for example, Gelman et al (1995). In MCMC, a sequence of draws is generated which follows a Markov chain, moving stochastically through the parameter space. Under certain conditions, the Markov chain will converge to its stationary distribution regardless of how the chain is initialized (e.g., Tierney, 1994). Successive draws from the Markov chain are then dependent but identically distributed draws from the stationary distribution. Our goal is to construct a Markov chain with the appropriate stationary distribution, namely the posterior distribution, $p(\theta \mid y)$. If the dependence in the chain is sufficiently weak, then $M^{-1} \sum_{t=1}^{M} h(\theta^{(t)}) \to \mathbb{E}\left(h(\theta) \mid y\right)$ almost surely as $M \to \infty$.

One method of generating this sequence of draws is to use the method of successive substitution sampling, more commonly referred to as Gibbs sampling. This method was

introduced by Geman and Geman (1984) and was generalized by Gelfand and Smith (1990). The idea of Gibbs sampling is that $\theta^{(t-1)}$ is updated to $\theta^{(t)}$ by updating each of J subvectors of θ in turn. We do this by obtaining a draw from $p(\theta_j \mid y, \theta_{-j}^{(t-1)})$, where θ_{-j} refers to the elements of θ not contained in θ_j. That is,

$$\theta_{-j}^{(t-1)} = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots \theta_J^{(t-1)}).$$

Thus, to implement the Gibbs sampler, the conditional posterior distribution, $p(\theta \mid y, \theta_{-j})$ is needed for each subvector. Not only do we need to know the form of the conditional posterior distributions, but we must have available some method of obtaining a draw from each of them.

Gibbs sampling cannot be applied to every model. It is not always possible to specify prior distributions and to choose subvectors of θ in such a way that the conditional posterior distributions are standard distributions. In this analysis, however, it is possible to specify a reasonable model for which Gibbs sampling is appropriate. Sections 6.2 and 6.5 describe the hierarchical model. Section 6.6 presents the conditional posterior distributions needed for implementing a Gibbs sampler.

6.1.4 Assessing convergence of the Gibbs sampler

A common issue when using MCMC methods is the length of the Markov chain needed to ensure that the draws are from the stationary distribution, regardless of the starting values used. To diffuse the issue of starting values, we will discard the first half of the draws from the Gibbs sampler. To determine how many draws are needed to be confident of convergence, we will use a method introduced by Gelman and Rubin (1992). This method involves the use of multiple Markov chains to assess the convergence of the chains to the stationary distribution. A statistic is used which compares the within chain variation to the between chain variation. If these two sources of variation are approximately the same, then we feel confident that the chains have converged to the stationary distribution.

Let k be the number of independent Markov chains used and let M denote half the length of each chain. Denote the draws of any scalar parameter by θ_{ij} for $i = 1, \ldots, M$

and $j = 1, \ldots, k$. Then the between and within chain variation are given by

$$B = \frac{M}{k-1} \sum_{j=1}^{k} (\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2$$
$$W = \frac{1}{k} \sum_{j=1}^{k} s_j^2,$$

where

$$\bar{\theta}_{\cdot j} = \frac{1}{M} \sum_{i=1}^{M} \theta_{ij},$$
$$\bar{\theta}_{\cdot\cdot} = \frac{1}{k} \sum_{j=1}^{k} \bar{\theta}_{\cdot j}, \text{and}$$
$$s_j^2 = \frac{1}{M-1} \sum_{i=1}^{M} (\theta_{ij} - \bar{\theta}_{\cdot j})^2.$$

The scale reduction statistic is defined by

$$\sqrt{\hat{R}} = \left[W^{-1} \left(\frac{M-1}{M} W + \frac{1}{M} B \right) \right]^{\frac{1}{2}}. \tag{6.1}$$

The value of this statistic goes to one as $M \to \infty$ if the starting distribution is overdispersed or if the starting distribution is exactly the stationary distribution of the Markov chain.

6.1.5 Model diagnostics

To assess the fit of the model, we will use posterior predictive assessment as introduced by Gelman et al (1996). For each draw, $\boldsymbol{\theta}^{(t)}$, from the posterior distribution, a replicate sample is generated. If the replicate data are similar to the original sample data, we will conclude that the fit of the model is adequate. Similarity is a vague issue, but from knowledge of the data structure, some useful measures of similarity can be constructed.

For each draw, $\boldsymbol{\theta}^{(t)}$, a replicate sample, $\boldsymbol{y}^{(t)}$, is constructed from the posterior predictive distribution, $p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}^{(t)})$. Any statistic that can be calculated for the original data can also be calculated for each replicate sample. In this way, we obtain draws from the distribution of the statistic under the specified model. The value of the statistic calculated from the data can be compared to the distribution of the statistic obtained from the replicate samples in the same way a statistic is compared to a reference distribution

in classical analysis. If the data statistic can be believed to be a draw from the posterior reference distribution, then the specified model appears to produce replicate data which is similar to the original data. If not, this may suggest inadequacies of the model.

Appropriateness of the prior and hyperprior distributions may be assesed through a similar approach. However, it may be necessary to use "pseudo-statistics" which depend on the values of some of the parameters of the model and thus are not truly statistics. Gelman et al. (1996) call these parameter-dependent statistics *discrepancies*. At each step of the Gibbs sampler, the discrepancy is calculated for the orginal data and for each replicate sample. A scatter plot of these pairs should cluster around the 45° line through the origin if the replicate data are similar to the original data.

6.2 Data model

6.2.1 Overview

The remainder of this chapter is devoted to development of a hierarchical model for obtaining quantile profiles for the soil texture data presented in Chapter 3. We will use the transformed data as in the imputation approach. The log-ratio transformation was defined in (3.1). This transformation allows us to work with two-dimensional vectors which can take on any values in $\mathbb{R}^2$ and to model a dependence structure among the components. Recall that the log-ratio transformed field and laboratory texture observations for horizon h of site g are denoted $\boldsymbol{f}_{gh} = (f_{gh,1}, f_{gh,2})'$ and $\boldsymbol{l}_{gh} = (l_{gh,1}, l_{gh,2})'$, respectively.

We assume that $\boldsymbol{l}_{gh}$ has a distribution which depends on the master horizon designation, m_{gh}, through its mean and variance. We also assume a site-specific random effect which allows non-zero correlation among $\{\boldsymbol{l}_{gh}\}_{h=1,\ldots,H_g}$ for a particular site g. The conditional distribution of $\boldsymbol{f}_{gh}$ given $\boldsymbol{l}_{gh}$ has a mean which depends on $\boldsymbol{l}_{gh}$.

The master horizon designation, order and depth of horizons at a site will be referred to as the horizon profile. Horizon profiles are modeled as following a Markov chain. The transition probabilities of the Markov chain are unknown parameters which are constant

from the surface to the bottom of the profile (48 inches). However, in Section 6.9, a generalization of the model which allows these transition probabilities to change over depth is investigated.

For an observed horizon at a site in $\mathcal{S} \cup \mathcal{F}$, $\boldsymbol{l}_{gh}$ is missing. This set of missing values is considered as part of the parameter vector. Thus, the marginal posterior distribution of the parameters of interest is averaged over the distribution of these missing values. We can view this averaging process as imputing values for $\{\boldsymbol{l}_{gh}\}_{g \in \mathcal{S} \cup \mathcal{F}, h=1,\ldots,H_g}$. This is similar to the calibration step of the imputation approach. However, in the Bayesian approach, there is nothing comparable to the imputation step of the imputation approach. That is, we are not imputing full profiles for sites in $\mathcal{S}$.

Quantile profiles for laboratory texture determinations are obtained from a mixture of the posterior predictive distributions of $\boldsymbol{l}_{gh}$ given m_{gh} for $m_{gh} = A, B, C$. The mixing coefficients come from the horizon profile model and are not constant across inches. The coefficients correspond to the probability that inch i falls in an A, B, or C horizon.

We use the notation $\theta \sim$ to describe the distribution of θ. The distributions we refer to most often are the normal, the inverse-Wishart and the Dirichlet. These distributions will be abbreviated $\mathbb{N}$, $\mathbb{IW}$ and $\mathbb{D}$, respectively. Note that the inverse-Wishart distribution is parameterized such that if a $k \times k$ matrix $\boldsymbol{M} \sim \mathbb{IW}_{\nu}\left(\boldsymbol{S}^{-1}\right)$, then $\mathbb{E}(\boldsymbol{M}) = (\nu - k - 1)^{-1}\boldsymbol{S}$.

6.2.2 Field and laboratory measurements

We assume the transformed field measurements are normally distributed with homoskedastic errors and a mean that depends on the corresponding transformed laboratory measurements. This differs from the calibration model used in the imputation approach. The main objective of the calibration model was to predict $\boldsymbol{l}_{gh}$ which was not a fixed quantity. Here, we are specifying the distribution of $\boldsymbol{f}_{gh}$ conditioned on the value of $\boldsymbol{l}_{gh}$. That is, for this level of the model, the values of $\boldsymbol{l}_{gh}$ are fixed.

We write

$$\begin{aligned} \boldsymbol{f}_{gh} &= \begin{pmatrix} \psi_{01} \\ \psi_{02} \end{pmatrix} + \begin{pmatrix} \psi_{11} & 0 \\ 0 & \psi_{22} \end{pmatrix} \begin{pmatrix} l_{gh,1} \\ l_{gh,2} \end{pmatrix} + \begin{pmatrix} \omega_{gh,1} \\ \omega_{gh,2} \end{pmatrix} \\ &= \boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 \boldsymbol{l}_{gh} + \boldsymbol{\omega}_{gh}, \end{aligned} \tag{6.2}$$

where

$$\{\boldsymbol{\omega}_{gh}\} \,|\, \boldsymbol{\Sigma}_\omega \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}_\omega),$$

for a positive-definite matrix $\boldsymbol{\Sigma}_\omega$ and

$$\boldsymbol{\Sigma}_\omega \sim \mathbb{IW}_{\nu_\omega}\left(S_\omega^{-1}\right).$$

Let $\boldsymbol{\psi} = (\psi_{01}, \psi_{11}, \psi_{02}, \psi_{22})'$. We assume

$$\boldsymbol{\psi} \sim \mathbb{N}(\boldsymbol{b}, \boldsymbol{V}_\psi),$$

Define

$$\boldsymbol{L}_{gh} = \begin{pmatrix} 1 & l_{1gh} & 0 & 0 \\ 0 & 0 & 1 & l_{2gh} \end{pmatrix}.$$

Then equation (6.2) can be written as

$$\boldsymbol{f}_{gh} = \boldsymbol{L}_{gh}\boldsymbol{\psi} + \boldsymbol{\omega}_{gh}.$$

Both forms of the notation will be useful. Note that the components of $\boldsymbol{f}_{gh}$ are not restricted to be independent since the off-diagonal elements of $\boldsymbol{\Sigma}_\omega$ may be non-zero.

The transformed laboratory measurements are assumed to be normally distributed with a mean and variance which depend on the master horizon designation. The model also includes a random effect for each site, $\boldsymbol{\alpha}_g$. We assume the $\{\boldsymbol{\alpha}_g\}$ are independent and identically distributed (iid) with

$$\boldsymbol{\alpha}_g \,|\, \boldsymbol{\Sigma}_\alpha \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}_\alpha),$$

for a positive-definite matrix $\boldsymbol{\Sigma}_\alpha$, where

$$\boldsymbol{\Sigma}_\alpha \sim \mathbb{IW}_{\nu_\alpha}\left(\boldsymbol{S}_\alpha^{-1}\right).$$

We assume

$$l_{gh} = \mu_{m_{gh}} + \alpha_g + \zeta_{gh},$$

where

$$\mu_A \sim \mathbb{N}(A, V_A), \quad \mu_B \sim \mathbb{N}(B, V_B), \quad \mu_C \sim \mathbb{N}(C, V_C),$$

the $\{\zeta_{gh}\}$ are distributed independently with

$$\zeta_{gh} \,|\, \Sigma_A, \Sigma_B, \Sigma_C \sim \mathbb{N}(0, \Sigma_{m_{gh}}),$$

for positive-definite matrices Σ_A, Σ_B and Σ_C, such that

$$\Sigma_A \sim \mathbb{IW}_{\nu_A}(S_A^{-1}), \quad \Sigma_B \sim \mathbb{IW}_{\nu_B}(S_B^{-1}) \quad \text{and} \quad \Sigma_C \sim \mathbb{IW}_{\nu_C}(S_C^{-1}).$$

6.2.3 Horizon transitions

The distribution of the laboratory measurements depends on the horizon profile at the site. Each profile is assumed to begin with an A horizon. Let the distribution of the horizon profile be described by a Markov chain which evolves across inches with six possible states and a transition probability matrix $\mathbf{\Delta}$. States 1 through 6 correspond to continuing an A horizon, beginning a new A horizon, continuing a B horizon, beginning a new B horizon, continuing a C horizon, and beginning a new C horizon, respectively. Define $\delta(j,k)$ to be the (j,k)th element of $\mathbf{\Delta}$. That is, $\delta(j,k) = \mathbb{P}(T_{gi} = k \mid T_{g,i-1} = j)$ for $i = 2, \ldots, 48$. Note that $\mathbf{\Delta}$ has the form

$$\begin{pmatrix} (*) & (*) & 0 & (*) & 0 & (*) \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & (*) & (*) & (*) & 0 & (*) \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & (*) & 0 & (*) & (*) & (*) \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \tag{6.3}$$

where $(*)$ indicates an unknown transition probability. For example, $\delta(1,3) = 0$, means that the probability of continuing a B horizon in the current inch given that the previous

inch was a continuation of an A horizon is zero. On the other hand, $\delta(1,4)$ is not necessarily zero. This parameter represents the probability of beginning a B horizon in the current inch given that the previous inch was a continuation of an A horizon.

Note that the transition matrix $\boldsymbol{\Delta}$ is not indexed by inch. That is, we assume that the conditional transition probabilities are constant over the length of the profile. However, this does not mean that the probability of being in a particular state is constant across inches. The vector of probabilities for the six states for inch i for any site g is $\left[(0,1,0,0,0,0)\boldsymbol{\Delta}^{i-1}\right]'$, since the initial state (inch 1) of the Markov chain is always state 2. As $i \to \infty$, this vector approaches its stationary distribution, but we consider only the first 48 states of the chain.

We will also use the notation

$$\begin{aligned} \boldsymbol{\delta}_A &= (\delta(1,1), \delta(1,2), \delta(1,4), \delta(1,6)), \\ \boldsymbol{\delta}_B &= (\delta(3,2), \delta(3,3), \delta(3,4), \delta(3,6)) \\ \text{and} \quad \boldsymbol{\delta}_C &= (\delta(5,2), \delta(5,4), \delta(5,5), \delta(5,6)). \end{aligned}$$

The subscript of each $\boldsymbol{\delta}$ indicates that it contains the transition probabilities pertaining to transitioning *from* a continuation of that type of horizon. Note that these three vectors contain all of the unknown elements of $\boldsymbol{\Delta}$ and that the elements of each vector must sum to one. We assume that

$$\begin{aligned} \boldsymbol{\delta}_A &\sim \mathbb{D}(d_{A1}, d_{A2}, d_{A3}, d_{A4}), \\ \boldsymbol{\delta}_B &\sim \mathbb{D}(d_{B1}, d_{B2}, d_{B3}, d_{B4}) \\ \text{and} \quad \boldsymbol{\delta}_C &\sim \mathbb{D}(d_{C1}, d_{C2}, d_{C3}, d_{C4}). \end{aligned}$$

The value of the Markov chain for inch i at site g is denoted T_{gi}. The values of $\{T_{gi}\}_{g \in \mathcal{D}; i=1,\ldots,I_g}$ are observed. Note that I_g is random; it is likely to be around 6 to 12 for $g \in \mathcal{S}$ and should be 48 for all $g \in \mathcal{F} \cup \mathcal{L}$. However, this is not always the case. If a horizon ended "near" 48 inches, I_g is the lower boundary of the last horizon. When $I_g \neq 48$, we assume that I_{g+1} is the beginning of a new horizon. Note that this is a reasonable assumption for $g \in \mathcal{S}$ also.

6.3 Quantile profiles

Recall that the main objective of analyzing the soil texture data is the same as described in Chapter 3. We wish to describe the distribution of laboratory texture profiles using marginal quantile profiles. Under the model, the distribution of a transformed laboratory measurement at inch i is a mixture of normal distributions. This mixture distribution can be used to produce quantile profiles analogous to those produced by the imputation approach described in Chapter 5.

The coefficients in the normal mixture depend on $\boldsymbol{\Delta}$. Let $\boldsymbol{l}_{\cdot\cdot i}$ be a transformed laboratory measurement for inch i. The distribution is assumed to be the same for all sites, so that the subscripts g and h are both replaced by $\cdot$. We write the distribution of $\boldsymbol{l}_{\cdot\cdot i}$ as

$$\boldsymbol{l}_{\cdot\cdot i} \,|\, \cdot \sim \sum_{m\in\{A,B,C\}} \mathcal{M}_{mi} \times \mathbb{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_\alpha), \tag{6.4}$$

where

$$\begin{aligned} \mathcal{M}_{Ai} =& [\boldsymbol{\Delta}^{i-1}]_{21} + [\boldsymbol{\Delta}^{i-1}]_{22}, \\ \mathcal{M}_{Bi} =& [\boldsymbol{\Delta}^{i-1}]_{23} + [\boldsymbol{\Delta}^{i-1}]_{24} \\ \text{and} \quad \mathcal{M}_{Ci} =& [\boldsymbol{\Delta}^{i-1}]_{25} + [\boldsymbol{\Delta}^{i-1}]_{26} \end{aligned}$$

and $[\boldsymbol{\Delta}^{i-1}]_{jk}$ is the jkth element of the matrix $\boldsymbol{\Delta}^{i-1}$.

We wish to obtain marginal quantiles for each component of soil texture. Clay quantiles are used to demonstrate how this can be done. The distribution function of a laboratory clay measurement, $c_{1i}^{(l)}$, at the ith inch is

$$\begin{aligned} \mathbb{P}\left(c_{\cdot\cdot i,1}^{(l)} \le q\right) =& \mathbb{P}\left(\frac{\exp(l_{i,1})}{\exp(l_{i,1}) + \exp(l_{i,2}) + 1} \le q\right) \\ =& \mathbb{P}\left(l_{i,1} \le \log\left[\frac{q}{q-1}(\exp(l_{i,2}) + 1)\right]\right) \\ =& \sum_{m\in\{A,B,C\}} \mathcal{M}_{mi} \mathbb{E}\left[\Phi\left(\frac{\log\left[\frac{q}{q-1}(\exp(l) + 1)\right] - \mathbb{E}\left(l_{i,1} \,|\, l_{i,2} = l\right)}{\sqrt{\mathrm{Var}\left(l_{i,1} \,|\, l_{i,2} = l\right)}}\right)\right], \end{aligned} \tag{6.5}$$

where $l \sim \mathbb{N}(\mu_{m,2}, [\Sigma_m]_{22})$, $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable and

$$\mathbb{E}\left(l_{i,1} \,|\, l_{i,2} = l\right) = \mu_{m,1} + \frac{[\Sigma_m]_{12}}{[\Sigma_m]_{22}}(l - \mu_{m,2}),$$
$$\mathrm{Var}\left(l_{i,1} \,|\, l_{i,2} = l\right) = [\Sigma_m]_{11} - \frac{([\Sigma_m]_{12})^2}{[\Sigma_m]_{11}},$$

for i in master horizon m. Posterior marginal clay quantiles can be obtained by inverting this distribution. Thus the quantiles of interest can be expressed as a function of the parameters of the model.

6.4 Maximum likelihood estimation

The parameter vector is $\theta = (\psi, \Sigma_\omega, \mu_A, \mu_B, \mu_C, \Sigma_\alpha, \Sigma_A, \Sigma_B, \Sigma_C, \delta_A, \delta_B, \delta_C)$. The collection of available data is denoted

$$\mathcal{Z} = \left\{ \{f_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g}, \{l_{gh}\}_{g\in\mathcal{L};h=1,\ldots,H_g}, \{m_{gh}, d_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g} \right\}.$$

The values of I_g and $\{T_{gi}\}_{g\in\mathcal{D};i=1,\ldots,I_g}$ can be calculated from elements of $\mathcal{Z}$. The data model is

$$\begin{aligned} p(\mathcal{Z} \,|\, \theta) =& p(\{f_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g} \,|\, \{l_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g}, \psi, \Sigma_\omega) \\ & \times p(\{l_{gh}\}_{g\in\mathcal{L};h=1,\ldots,H_g} \,|\, \{m_{gh}\}_{g\in\mathcal{L};h=1,\ldots,H_g}, \mu_A, \mu_B, \mu_C, \Sigma_A, \Sigma_B, \Sigma_C, \{\alpha_g\}_{g\in\mathcal{L}}) \\ & \times p(\{T_{gi}\}_{g\in\mathcal{D};i=1,\ldots,I_g}, \{d_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g} \,|\, \delta_A, \delta_B, \delta_C). \end{aligned} \tag{6.6}$$

When viewed as a function of θ, (6.6) is called the likelihood. Maximum likelihood estimates (MLEs) are obtained by choosing $\theta = \hat{\theta}$ such that (6.6) is maximized. Note that the likelihood factors into a factor for the field and laboratory measurement models and a factor for the horizon profile model.

A special case of the data model is when $\alpha_g = 0$ for $g \in \mathcal{D}$. In this case, we can derive MLEs of the parameters in closed form. We will denote the MLE of θ by $\hat{\theta}$. For

the transition probabilities, we have that

$$\hat{\delta}(j,k) = \left(\sum_{g,i} \mathbb{I}\left(T_{g,i-1} = j\right) \right)^{-1} \left(\sum_{g,i} \sum_{k' \in \{2,4,6\}} \mathbb{I}\left(T_{g,i-1} = j, T_{gi} = k'\right) + \sum_{g} \mathbb{I}\left(T_{g,I_g} = j, I_g \neq 48\right) \frac{\sum_{g,i} \mathbb{I}\left(T_{g,i-1} = j, T_{gi} = k\right)}{\sum_{k' \in \{2,4,6\}} \mathbb{I}\left(T_{g,i-1} = j, T_{gi} = k'\right)} \right)$$

and

$$\hat{\delta}(j,j) = 1 - \sum_{k' \in \{2,4,6\}} \hat{\delta}(j,k')$$

for $j = 1,3,5$ and $k = 2,4,6$. Let $\hat{\mathbf{\Delta}}$ represent the transition matrix containing these MLEs. Then the MLEs of the coefficients of the mixture distribution are

$$\begin{aligned} \hat{\mathcal{M}}_{Ai} &= [\hat{\mathbf{\Delta}}^{i-1}]_{21} + [\hat{\mathbf{\Delta}}^{i-1}]_{22}, \\ \hat{\mathcal{M}}_{Bi} &= [\hat{\mathbf{\Delta}}^{i-1}]_{23} + [\hat{\mathbf{\Delta}}^{i-1}]_{24} \\ \text{and} \quad \hat{\mathcal{M}}_{Ci} &= [\hat{\mathbf{\Delta}}^{i-1}]_{25} + [\hat{\mathbf{\Delta}}^{i-1}]_{26}. \end{aligned}$$

Define the quantities

$$\begin{aligned} \bar{f}_{\mathcal{D},m} &= \frac{\sum_{g \in \mathcal{D}} \sum_{h=1}^{H_g} f_{gh}}{\sum_{g \in \mathcal{D}} \sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh} = m\right)}, \\ \bar{f}_{\mathcal{L},m} &= \frac{\sum_{g \in \mathcal{L}} \sum_{h=1}^{H_g} f_{gh}}{\sum_{g \in \mathcal{L}} \sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh} = m\right)}, \\ \bar{l}_{\mathcal{L},m} &= \frac{\sum_{g \in \mathcal{L}} \sum_{h=1}^{H_g} l_{gh}}{\sum_{g \in \mathcal{L}} \sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh} = m\right)}, \end{aligned}$$

$$S_{\mathcal{D},ff,m} = \left(\sum_{g\in\mathcal{D}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\right)^{-1} \sum_{g\in\mathcal{D}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\left(f_{gh}-\bar{f}_{\mathcal{D},m}\right)\left(f_{gh}-\bar{f}_{\mathcal{D},m}\right)',$$

$$S_{\mathcal{L},ff,m} = \left(\sum_{g\in\mathcal{L}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\right)^{-1} \sum_{g\in\mathcal{L}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\left(f_{gh}-\bar{f}_{\mathcal{L},m}\right)\left(f_{gh}-\bar{f}_{\mathcal{L},m}\right)',$$

$$S_{\mathcal{L},ll,m} = \left(\sum_{g\in\mathcal{L}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\right)^{-1} \sum_{g\in\mathcal{L}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\left(l_{gh}-\bar{l}_{\mathcal{L},m}\right)\left(l_{gh}-\bar{l}_{\mathcal{L},m}\right)',$$

$$S_{\mathcal{L},lf,m} = \left(\sum_{g\in\mathcal{L}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\right)^{-1} \sum_{g\in\mathcal{L}}\sum_{h=1}^{H_g} \mathbb{I}\left(m_{gh}=m\right)\left(l_{gh}-\bar{l}_{\mathcal{L},m}\right)\left(f_{gh}-\bar{f}_{\mathcal{L},m}\right)',$$

$$\text{and} \quad \hat{R}_m = S_{\mathcal{L},lf,m} S_{\mathcal{L},ff,m}^{-1}. \tag{6.7}$$

Then, using results from Anderson (1957), the MLEs for the parameters of the laboratory measurement model are

$$\begin{aligned}\hat{\mu}_m &= \bar{l}_{\mathcal{L},m} + \hat{R}_m(\bar{f}_{\mathcal{D},m} - \bar{f}_{\mathcal{L},m}),\\ \hat{\Sigma}_m &= S_{\mathcal{L},ll,m} + \hat{R}_m\left(S_{\mathcal{D},ff,m} - S_{\mathcal{L},ff,m}\right)\hat{R}_m'.\end{aligned}$$

The MLE of a quantile profile can be obtained by plugging 6.7 and 6.8 in to (6.5) and inverting the estimated distribution function. MLEs of the parameters of field measurement are not needed. However, evaluating the quality of these estimates depends on the asymptotic normality of a MLE. It is not clear how good this approximation would be for the sample sizes in the soil texture data. Also, the MLEs do not have closed forms if the site-specific random effect has non-zero variance. Further, future work on the soils project will involve small area estimation for soil map units. For these reasons, a Bayesian approach has considerable appeal.

6.5 Prior distributions

For Bayesian inference, we must specify a prior distribution for θ. The prior distribution of θ is assumed to be of the form

$$\begin{aligned}p(\theta) =& p(\psi)p(\Sigma_\omega)p(\Sigma_\alpha)p(\mu_A)p(\mu_B)p(\mu_C)p(\Sigma_A)p(\Sigma_B)p(\Sigma_C)\\ &\times p(\delta_A)p(\delta_B)p(\delta_C).\end{aligned} \tag{6.8}$$

Each of the factors of the prior distribution were given in Section 6.2 and are summarized in Table 6.1. For each factor, a distribution that is conjugate for the corresponding conditional posterior distribution was chosen. A conjugate prior results in a posterior distribution in the same family of distributions as the prior distribution.

The parameters of each prior distribution are chosen so that the distribution is vague. That is, the variance is very large relative to what is believable. A large variance represents our uncertainty in the value of the parameter. The location parameters for each prior distribution were chosen based on "reasonable" values according to the data. The values of the parameters of the prior distributions used in the analysis are given in Table 6.1.

6.6 Conditional posterior distributions and Gibbs sampler

If $I_g \neq 48$, we have some partial information about the value of T_{g,I_g+1}. That is, when $I_g \neq 48$, we assume that $T_{g,I_g+1} = 2, 4$ or 6. Similarly, when $\boldsymbol{f}_{gh}$ is observed, but $\boldsymbol{l}_{gh}$ is not, we have partial information about the value of $\boldsymbol{l}_{gh}$. For computational simplicity, we will augment the parameter vector with these "partially" observed values. Consider the full posterior distribution,

$$
\begin{aligned}
& p(\boldsymbol{\theta}, \{\boldsymbol{l}_{gh}\}_{g\in\mathcal{S}\cup\mathcal{F};h=1,\ldots,H_g}, \{\boldsymbol{\alpha}_g\}_{g\in\mathcal{D}}, \{T_{g,I_g+1}\}_{g\in\mathcal{D}:I_g<48} \mid \mathcal{Z}) \qquad (6.9)\\
\propto & p(\{\boldsymbol{f}_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g} \mid \{\boldsymbol{l}_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g}, \boldsymbol{\psi}, \boldsymbol{\Sigma}_\omega)\\
& \times p(\{\boldsymbol{l}_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g} \mid \{m_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g}, \boldsymbol{\mu}_A, \boldsymbol{\mu}_B, \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_C, \{\boldsymbol{\alpha}_g\}_{g\in\mathcal{D}})\\
& \times p(\{\boldsymbol{\alpha}_g\}_{g\in\mathcal{D}} \mid \boldsymbol{\Sigma}_\alpha)\\
& \times p(\{T_{gi}\}_{g\in\mathcal{D};i=1,\ldots,I_g}, \{d_{gh}\}_{g\in\mathcal{D};h=1,\ldots,H_g} \{T_{g,I_g+1}\}_{g\in\mathcal{D}:I_g<48} \mid \boldsymbol{\delta}_A, \boldsymbol{\delta}_B, \boldsymbol{\delta}_C)\\
& \times p(\boldsymbol{\theta}),
\end{aligned}
$$

where $p(\boldsymbol{\theta})$ is given in (6.8). In order to sample from the posterior distribution in (6.9), Gibbs sampling is used. The augmented parameter vector is divided into subvectors. For each subvector, the conditional posterior is needed in order to implement the Gibbs sampler. The subvectors have been chosen in such a way that each of the conditional

Table 6.1 Summary of prior distributions.

Portion of model	Hyper-parameter	Distribution
Field msmts.	ψ	$\mathbb{N}\left(\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 100 \end{bmatrix}\right)$
	Σ_ω	$\mathbb{IW}_4\left(\begin{bmatrix} 0.066 & 0.022 \\ 0.022 & 0.567 \end{bmatrix}^{-1}\right)$
Lab msmts.	Σ_α	$\mathbb{IW}_4\left(\begin{bmatrix} 0.213 & -0.120 \\ -0.120 & 0.672 \end{bmatrix}^{-1}\right)$
	μ_A	$\mathbb{N}\left(\begin{bmatrix} -0.216 \\ -2.260 \end{bmatrix}, \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}\right)$
	μ_B	$\mathbb{N}\left(\begin{bmatrix} -0.275 \\ -2.100 \end{bmatrix}, \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}\right)$
	μ_C	$\mathbb{N}\left(\begin{bmatrix} -0.782 \\ -1.760 \end{bmatrix}, \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}\right)$
	Σ_A	$\mathbb{IW}_4\left(\begin{bmatrix} 0.0346 & 0.0046 \\ 0.0046 & 0.319 \end{bmatrix}^{-1}\right)$
	Σ_B	$\mathbb{IW}_4\left(\begin{bmatrix} 0.117 & -0.037 \\ -0.037 & 0.183 \end{bmatrix}^{-1}\right)$
	Σ_C	$\mathbb{IW}_4\left(\begin{bmatrix} 0.141 & 0.034 \\ 0.034 & 0.548 \end{bmatrix}^{-1}\right)$
Horizon transitions	δ_A	$\mathbb{D}(1,1,1,1)$
	δ_B	$\mathbb{D}(1,1,1,1)$
	δ_C	$\mathbb{D}(1,1,1,1)$

posterior distributions is a recognizable distribution that is easily sampled.

In order to implement a Gibbs sampling algorithm, the conditional posterior distribution for each subvector of θ is needed. We will use the notation $\theta \,|\, \cdot \sim$ to describe to conditional posterior distribution of a parameter θ given all other parameters and the data. The shorthand $\sum_{g,h}$ will be used for $\sum_{g\in\mathcal{D}} \sum_{h=1}^{H_g}$. The conditional posterior distributions are as follows.

(*i*) Regression coefficient for field measurement model

$$\psi \,|\, \cdot \sim \mathbb{N}\left(\left[\sum_{g,h} L'_{gh}\Sigma_\omega^{-1} L_{gh} + V_\psi^{-1} \right]^{-1} \left[\sum_{g,h} L'_{gh}\Sigma_\omega^{-1} f_{gh} + V_\psi^{-1} b \right], \left[\sum_{g,h} L'_{gh}\Sigma_\omega^{-1} L_{gh} + V_\psi^{-1} \right]^{-1} \right)$$

(*ii*) Variance of residuals in field measurement model

$$\Sigma_\omega \,|\, \cdot \sim \mathbb{IW}_{(\nu_\omega + \sum_{g\in\mathcal{D}} H_g)} \left(\left[\sum_{g,h} \left(f_{gh} - L_{gh}\psi\right)\left(f_{gh} - L_{gh}\psi\right)' + S_e \right]^{-1} \right).$$

(*iii*) Missing lab data for $g \in \mathcal{S} \cup \mathcal{F}$

$$l_{gh} \,|\, \cdot \sim \mathbb{N}\left(\left[\psi_1'\Sigma_\omega^{-1}\psi_1 + \Sigma_{m_{gh}}^{-1} \right]^{-1} \left[\psi_1'\Sigma_\omega^{-1}\left(f_{gh} - \psi_0\right) + \Sigma_{m_{gh}}^{-1}\left(\mu_{m_{gh}} + \alpha_g\right) \right], \left[\psi_1'\Sigma_\omega^{-1}\psi_1 + \Sigma_{m_{gh}}^{-1} \right]^{-1} \right).$$

(*iv*) Means for laboratory measurement model

$$\mu_A \,|\, \cdot \sim \mathbb{N}\left(\mathcal{V}_A^{-1} \left[\Sigma_A^{-1} \sum_{g,h} \mathbb{I}\left(m_{gh} = B\right)\left(l_{gh} - \alpha_g\right) + V_A^{-1} A \right], \mathcal{V}_A^{-1} \right)$$

$$\mu_B \,|\, \cdot \sim \mathbb{N}\left(\mathcal{V}_B^{-1} \left[\Sigma_B^{-1} \sum_{g,h} \mathbb{I}\left(m_{gh} = B\right)\left(l_{gh} - \alpha_g\right) + V_B^{-1} B \right], \mathcal{V}_B^{-1} \right)$$

$$\mu_C \,|\, \cdot \sim \mathbb{N}\left(\mathcal{V}_C^{-1} \left[\Sigma_C^{-1} \sum_{g,h} \mathbb{I}\left(m_{gh} = C\right)\left(l_{gh} - \alpha_g\right) + V_C^{-1} C \right], \mathcal{V}_C^{-1} \right),$$

where $\mathcal{V}_m = \Sigma_m^{-1} \sum_{g,h} \mathbb{I}\left(m_{gh} = m\right) + V_m^{-1}$ for $m = A, B, C$.

(*v*) Site-specific random effect for laboratory measurement model

$$\boldsymbol{\alpha}_g \mid \cdot \sim \mathbb{N}\left(\mathcal{V}_\alpha^{-1}\left[\sum_{h=1}^{H_g}\boldsymbol{\Sigma}_{m_{gh}}^{-1}\left(\boldsymbol{l}_{gh}-\boldsymbol{\mu}_{m_{gh}}\right)\right],\mathcal{V}_\alpha^{-1}\right),$$

where $\mathcal{V}_\alpha = \sum_{h=1}^{H_g}\boldsymbol{\Sigma}_{m_{gh}}^{-1}+\boldsymbol{\Sigma}_\alpha^{-1}$.

(*vi*) Variance of random effects for laboratory measurement model

$$\boldsymbol{\Sigma}_\alpha \mid \cdot \sim \mathbb{IW}_{(\nu_\alpha+|\mathcal{D}|)}\left(\left[\sum_{g\in\mathcal{D}}\boldsymbol{\alpha}_g\boldsymbol{\alpha}_g'+\boldsymbol{S}_\alpha\right]^{-1}\right)$$

(*vii*) Variance of residuals for laboratory measurement model

$$\boldsymbol{\Sigma}_A \mid \cdot \sim \mathbb{IW}_{(\nu_A+\sum_{g,h}\mathbb{I}(m_{gh}=A))}\left(\left[\sum_{g,h}\mathbb{I}\left(m_{gh}=A\right)\left(\boldsymbol{l}_{gh}-\boldsymbol{\mu}_A-\boldsymbol{\alpha}_g\right)\left(\boldsymbol{l}_{gh}-\boldsymbol{\mu}_A-\boldsymbol{\alpha}_g\right)'+\boldsymbol{S}_A\right]^{-1}\right)$$

$$\boldsymbol{\Sigma}_B \mid \cdot \sim \mathbb{IW}_{(\nu_B+\sum_{g,h}\mathbb{I}(m_{gh}=B))}\left(\left[\sum_{g,h}\mathbb{I}\left(m_{gh}=B\right)\left(\boldsymbol{l}_{gh}-\boldsymbol{\mu}_B-\boldsymbol{\alpha}_g\right)\left(\boldsymbol{l}_{gh}-\boldsymbol{\mu}_B-\boldsymbol{\alpha}_g\right)'+\boldsymbol{S}_B\right]^{-1}\right)$$

$$\boldsymbol{\Sigma}_C \mid \cdot \sim \mathbb{IW}_{(\nu_C+\sum_{g,h}\mathbb{I}(m_{gh}=B))}\left(\left[\sum_{g,h}\mathbb{I}\left(m_{gh}=B\right)\left(\boldsymbol{l}_{gh}-\boldsymbol{\mu}_C-\boldsymbol{\alpha}_g\right)\left(\boldsymbol{l}_{gh}-\boldsymbol{\mu}_C-\boldsymbol{\alpha}_g\right)'+\boldsymbol{S}_C\right]^{-1}\right)$$

(*viii*) State of the Markov chain inch I_{g+1} if $I_g \neq 48$

$$\mathbb{P}\left(T_{g,I_g+1}=k\right)=\frac{\delta(j,k)}{\delta(j,2)+\delta(j,4)+\delta(j,6)}, \tag{6.10}$$

if $T_{g,I_g}=j$ for $j=1,3,5$ and $k=2,4,6$.

(*ix*) Transition probabilities for horizon profile model

$$\delta_A \mid \cdot \sim \mathbb{D}\Big(d_{A1} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 1, T_{g,i-1} = 1\right), d_{A2} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 2, T_{g,i-1} = 1\right),$$
$$d_{A3} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 4, T_{g,i-1} = 1\right), d_{A4} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 6, T_{g,i-1} = 1\right) \Big)$$
$$\delta_B \mid \cdot \sim \mathbb{D}\Big(d_{B1} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 2, T_{g,i-1} = 3\right), d_{B2} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 3, T_{g,i-1} = 3\right),$$
$$d_{B3} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 4, T_{g,i-1} = 3\right), d_{B4} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 6, T_{g,i-1} = 3\right) \Big)$$
$$\delta_C \mid \cdot \sim \mathbb{D}\Big(d_{C1} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 2, T_{g,i-1} = 5\right), d_{C2} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 4, T_{g,i-1} = 5\right),$$
$$d_{C3} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 5, T_{g,i-1} = 5\right), d_{C4} + \sum_{g,i} \mathbb{I}\left(T_{gi} = 6, T_{g,i-1} = 5\right) \Big),$$

where $\sum_{g,i}$ means $\sum_{g \in \mathcal{D}} \sum_{i=2}^{\min(48, I_g+1)}$.

6.7 Analysis results for the soil texture data

We use Gelman and Rubin's (1992) potential scale reduction statistic to assess convergence of the Markov Chain for the soil texture data. Five parallel chains were simulated. Starting values were drawn from the prior distributions. The values of the statistic for most scalar parameters indicate that 2000 iterations are sufficient for this Gibbs sampler to converge.

Table 6.2 contains 95% posterior intervals for some of the parameters. Under this model, $\boldsymbol{f}$ does not appear to be well calibrated to $\boldsymbol{l}$ since the posterior intervals for the slopes (ψ_{11}, ψ_{22}) do not include 1.0. Also, posterior intervals of the diagonal elements of $\boldsymbol{\Sigma}_\omega$ show that the measurement error in $\boldsymbol{f}$ is not very small relative to the variability in $\boldsymbol{l}$.

The posterior distribution for the means of the laboratory model, $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\mu}_C$ are shown in Figures 6.1 and 6.2. The posterior intervals for the off-diagonal elements of $\boldsymbol{\Sigma}_\omega$, $\boldsymbol{\Sigma}_A$, $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_C$ include zero, indicating that there is little or no covariance between the two components. However, this is not true for $\boldsymbol{\Sigma}_\alpha$. Figures 6.3 and 6.4

Table 6.2 Posterior intervals for some parameters.

Parameters	Interval
ψ	$\begin{pmatrix} (-0.2367, -0.1527) \\ (0.6325, 0.7789) \\ (-1.7004, -1.3301) \\ (.3430, .4824) \end{pmatrix}$
Σ_ω	$\begin{pmatrix} (0.0413, 0.0606) & (-0.0108, 0.0253) \\ (-0.0108, 0.0253) & (0.2659, 0.3726) \end{pmatrix}$
μ_A	$\begin{pmatrix} (-0.1482, 0.003) \\ (-2.7536, -2.3147) \end{pmatrix}$
μ_B	$\begin{pmatrix} (-0.2887, -0.6906) \\ (-2.5331, -2.1519) \end{pmatrix}$
μ_C	$\begin{pmatrix} (-0.4836, -0.1394) \\ (-2.6403, -2.1062) \end{pmatrix}$
Σ_A	$\begin{pmatrix} (0.0171, 0.0344) & (-0.0069, 0.0354) \\ (-0.0069, 0.0354) & (0.1961, 0.4063) \end{pmatrix}$
Σ_B	$\begin{pmatrix} (0.0864, 0.1876) & (-0.0887, 0.162) \\ (-0.0887, 0.162) & (0.1331, 0.3944) \end{pmatrix}$
Σ_C	$\begin{pmatrix} (0.1215, 0.3645) & (-0.1859, 0.1736) \\ (-0.1859, 0.1736) & (0.3553, 1.3886) \end{pmatrix}$
δ_A	$\begin{pmatrix} (0.8773, 0.9087) \\ (0.0676, 0.0965) \\ (0.0172, 0.0331) \\ (0.0028, 0.0110) \end{pmatrix}$
δ_B	$\begin{pmatrix} (0.0024, 0.0169) \\ (0.8676, 0.9186) \\ (0.0495, 0.0916) \\ (0.0240, 0.0558) \end{pmatrix}$
δ_C	$\begin{pmatrix} (0.0139, 0.0561) \\ (0.0001, 0.0147) \\ (0.8860, 0.9483) \\ (0.0325, 0.0888) \end{pmatrix}$
Σ_α	$\begin{pmatrix} (0.1732, 0.3497) & (-0.3035, -0.0556) \\ (-0.3035, -0.0556) & (0.5893, 1.47) \end{pmatrix}$

show histograms for the $(1,1)$ element and the $(2,2)$ element, respectively, of each of the variance components. Note that the elements of Σ_α are at least as large as the elements of Σ_A, Σ_B and Σ_C. Thus, we conclude that there is a site-specific random effect in the data.

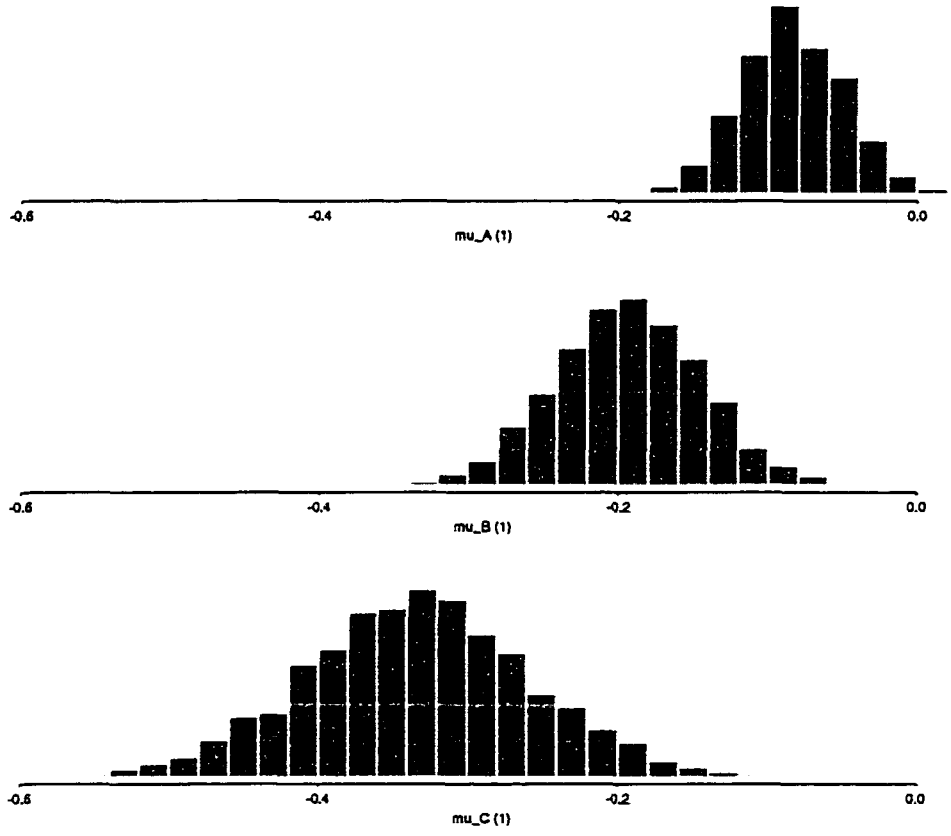

Figure 6.1 Posterior distributions of mean of the first component of $\boldsymbol{l}$.

Figure 6.5 shows the probability of being in each master horizon as a function of depth. In general, this plot follows an expected pattern. The probability of finding an A horizon decreases steadily from the top of the profile to the bottom. The probability of a B horizon rises more sharply than that of a C horizon for the first part of the profile and then begins to level off. However, even at the bottom of the profile, the probability of finding an A horizon is greater than either of the other two. This is not reasonable

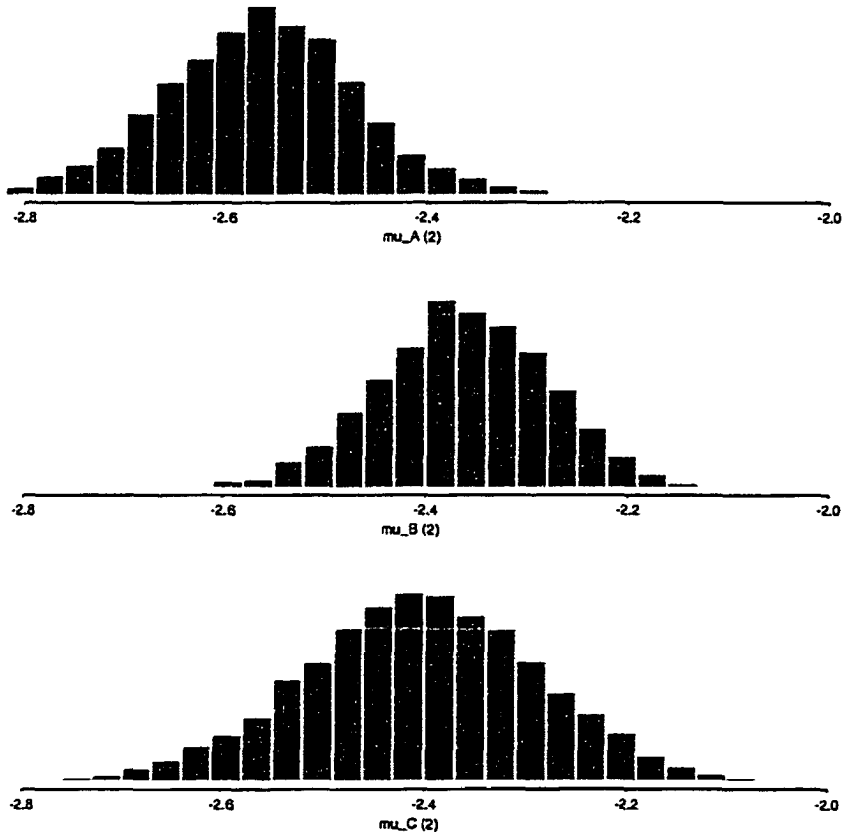

Figure 6.2 Posterior distributions of mean of the second component of l.

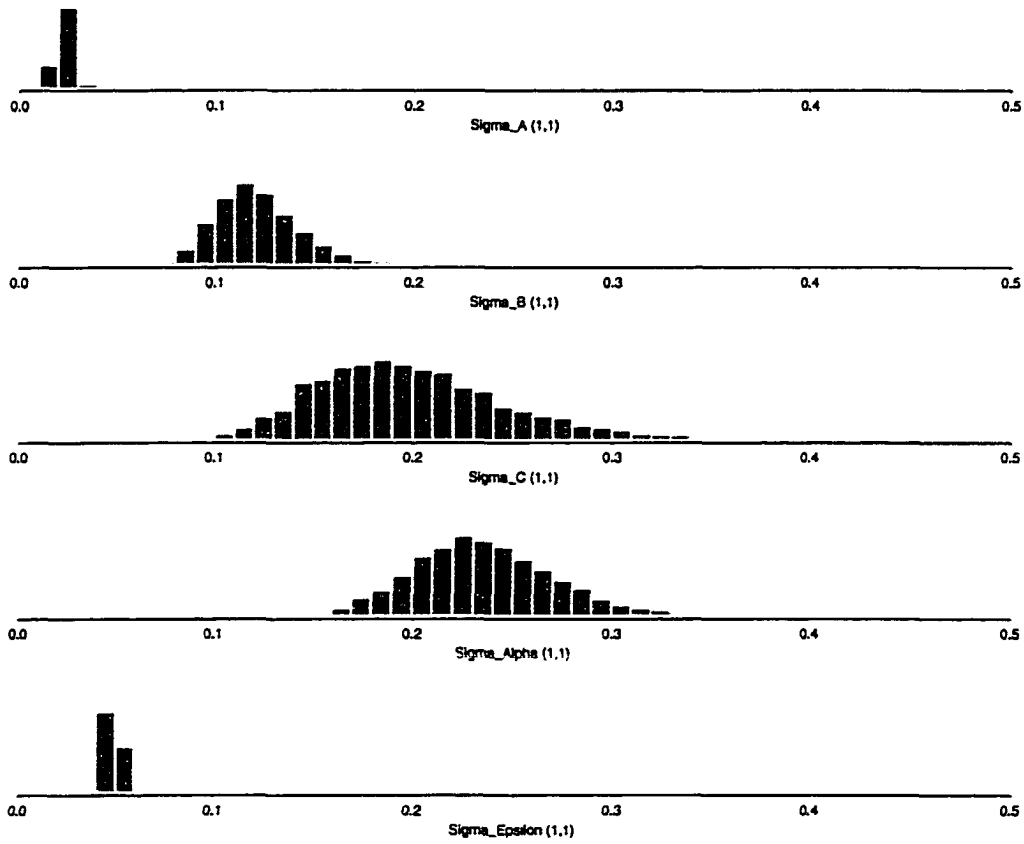

Figure 6.3 Posterior distributions of $(1,1)$ element of each variance component.

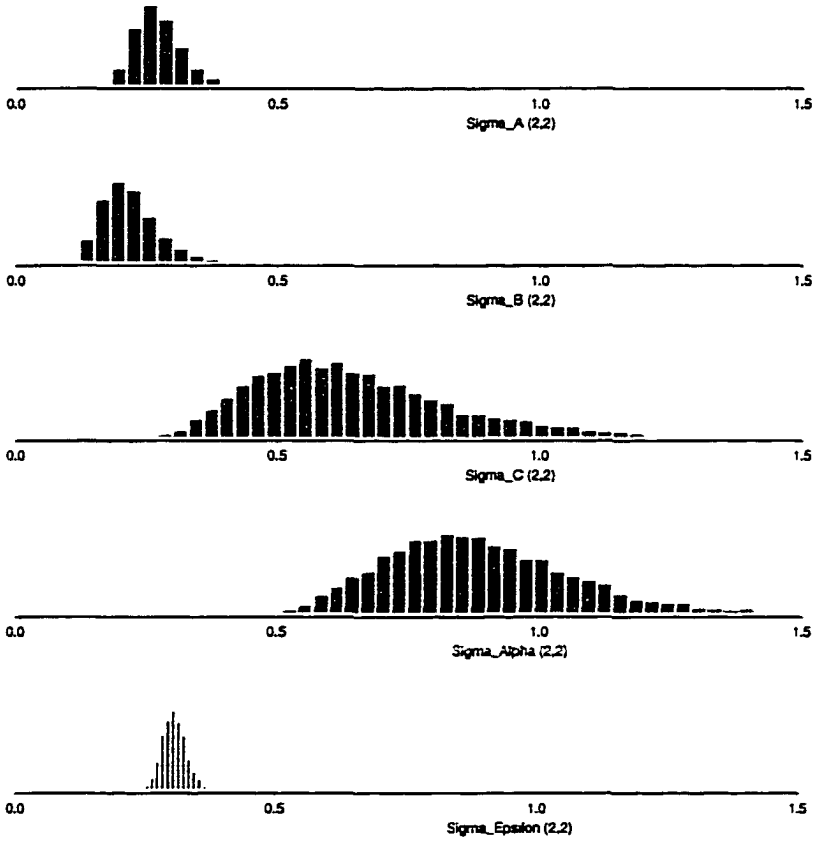

Figure 6.4 Posterior distributions of $(2,2)$ element of each variance component.

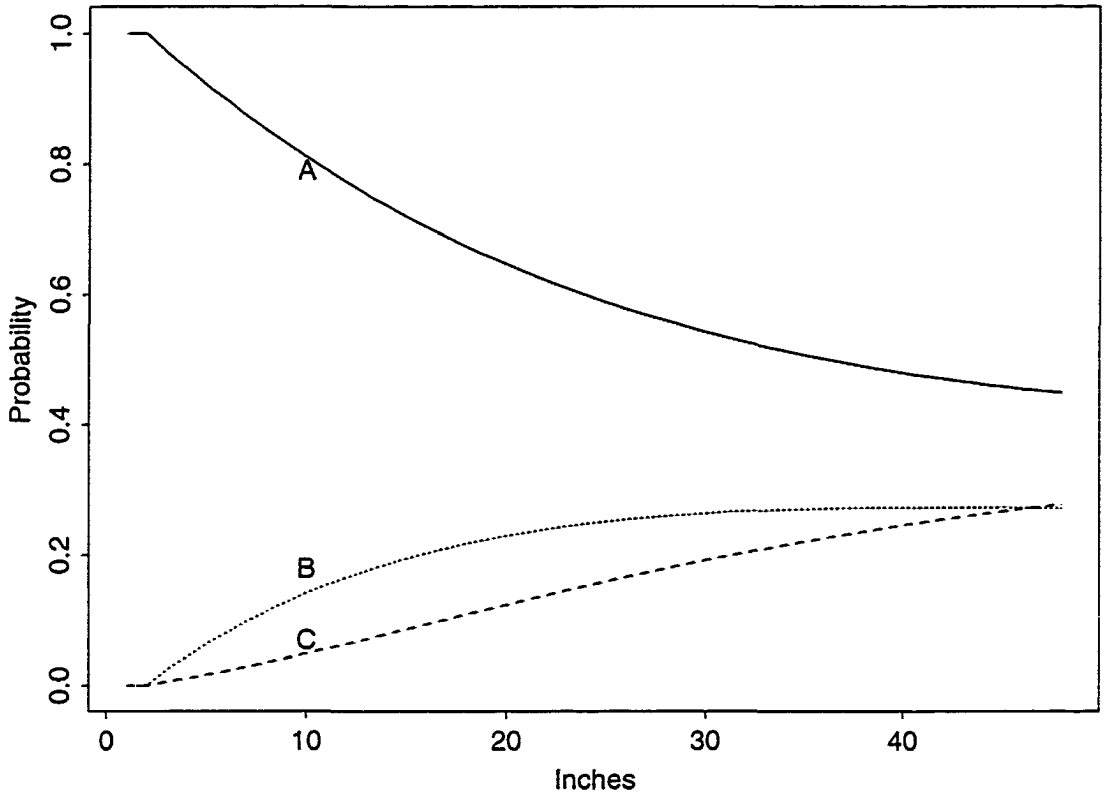

Figure 6.5 Posterior profiles for the probability of each master horizon designation occurring as a function of depth.

based on other information about horizon profiles in these soils. We will also see in the model diagnostics in Section 6.8 that this part of the model does not appear to fit well. An improvement to the horizon profile model is suggested in Section 6.9.

Figure 6.6 shows the observed data and estimated clay quantile profiles computed with parameters set equal to their posterior means. They are much smoother than those produced by the imputation approach. This is a product of the model, but is a desired feature of the estimates, as in the imputation approach.

We compare the observed data to the estimated quantiles by calculating the relative frequency of data points within several intervals across inches: 55% of the data points

are in the interquartile range; 76% of the data points fall between the estimated .10 and .90 quantiles; 87% of the data points fall between the estimated .05 and .95 quantiles; and 98% of the data points fall between the estimated .01 and .99 quantiles. These coverages are better than those of the imputation approach. However, we note that one-sided coverage is not as good. For example, there are more data points between the estimated .01 and .05 quantiles than between the estimated .95 and .99 quantiles.

This pattern seems to indicate the the true distribution is more left-skewed than this set of estimates. The estimated distribution looks fairly symmetric, in contrast to the left-skewed distribution estimated by the imputation approach, which appeared too peaked. Thus both approaches demonstrate some lack of fit.

Figure 6.7 shows boxplots of the posterior distributions for the .25, .50 and .75 quantiles (bottom to top) for each inch. These indicate that the .75 quantile is more variable than the median and the .25 quantile. This is different from the results in the imputation approach. Figure 6.8 shows the estimated standard deviations for the three quantiles from the two approach. The difference in ordering of the standard deviation profiles stems from the shape of the estimated distributions that each approach is producing.

6.8 Model checking

6.8.1 Field and laboratory measurement model

Figure 6.9 contains scatter plots of the two transformed components of field and laboratory measurements. The solid line in two of the plots represents the regression line using the posterior mean of ψ. These two regressions are the field measurement model. Figure 6.10 shows residual plots from the regression model using posterior means. The bottom right plot shows some possible heteroskedasticity in the residuals. Figure 6.11 contains normal probability plots for each component of $\boldsymbol{l}_{gh}$ for each master horizon designation. While there may be minor departures from the model assumptions for the field and laboratory measurement models, the assumptions of linearity and normality appear to be reasonable.

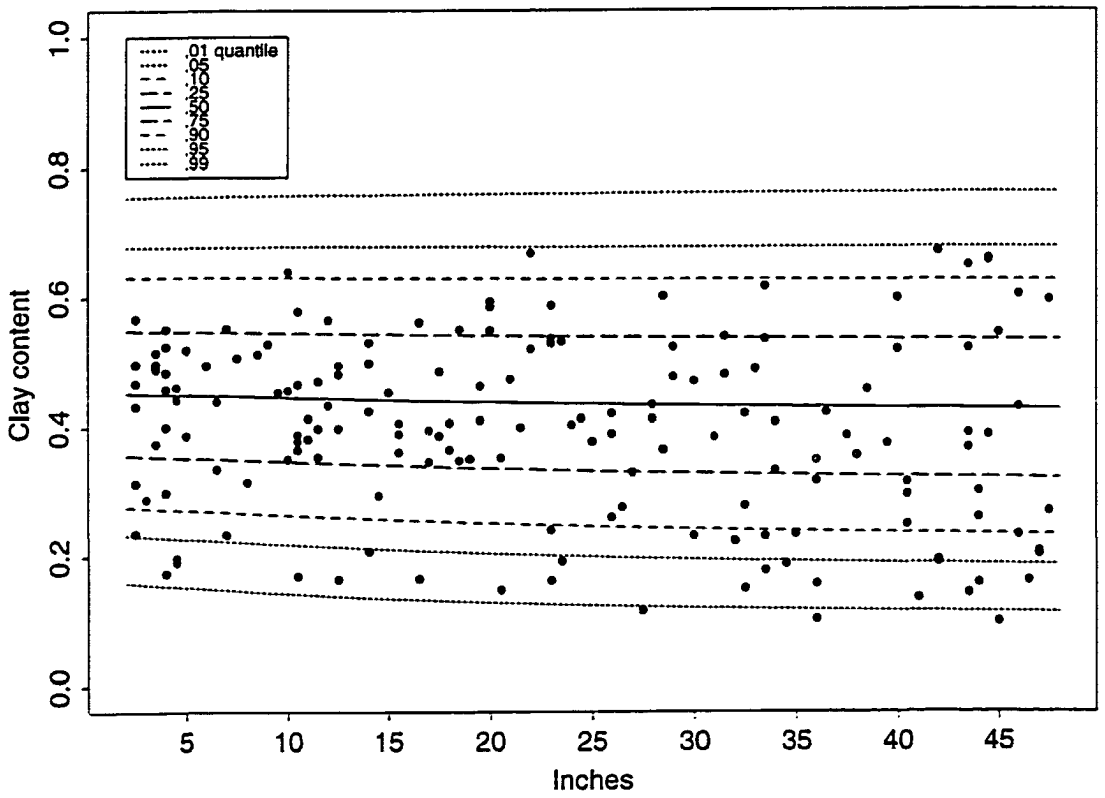

Figure 6.6 Posterior profiles of quantiles of the distribution of clay for Old Alluvium soils. The horizontal axis represents inches; the vertical axis represents clay content. The solid, dashed and dotted lines are quantiles as indicated in the legend. Dots on the graph represent laboratory determinations of clay content for each horizon at each site $g \in \mathcal{L}$. Each dot is plotted at the midpoint of the horizon.

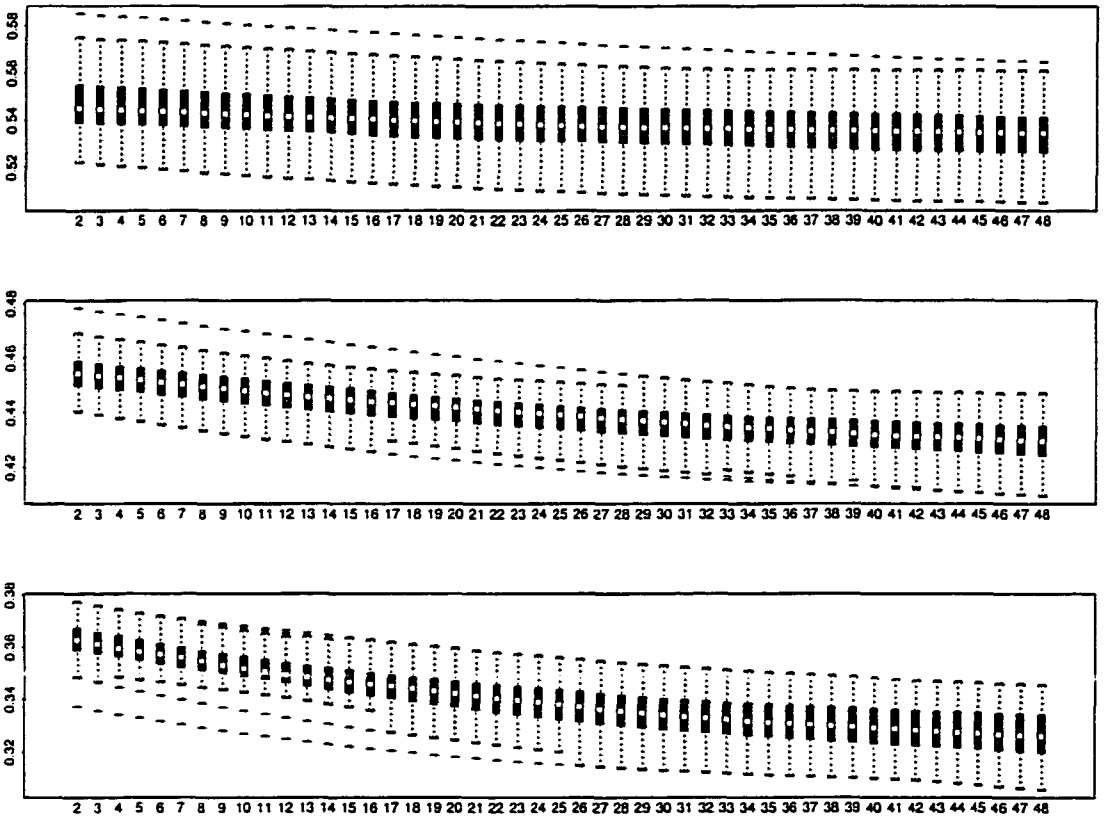

Figure 6.7 Posterior distributions of quantiles of the distribution of clay. Top plot is the upper quartile; middle plot is the median; bottom plot is the lower quartile.

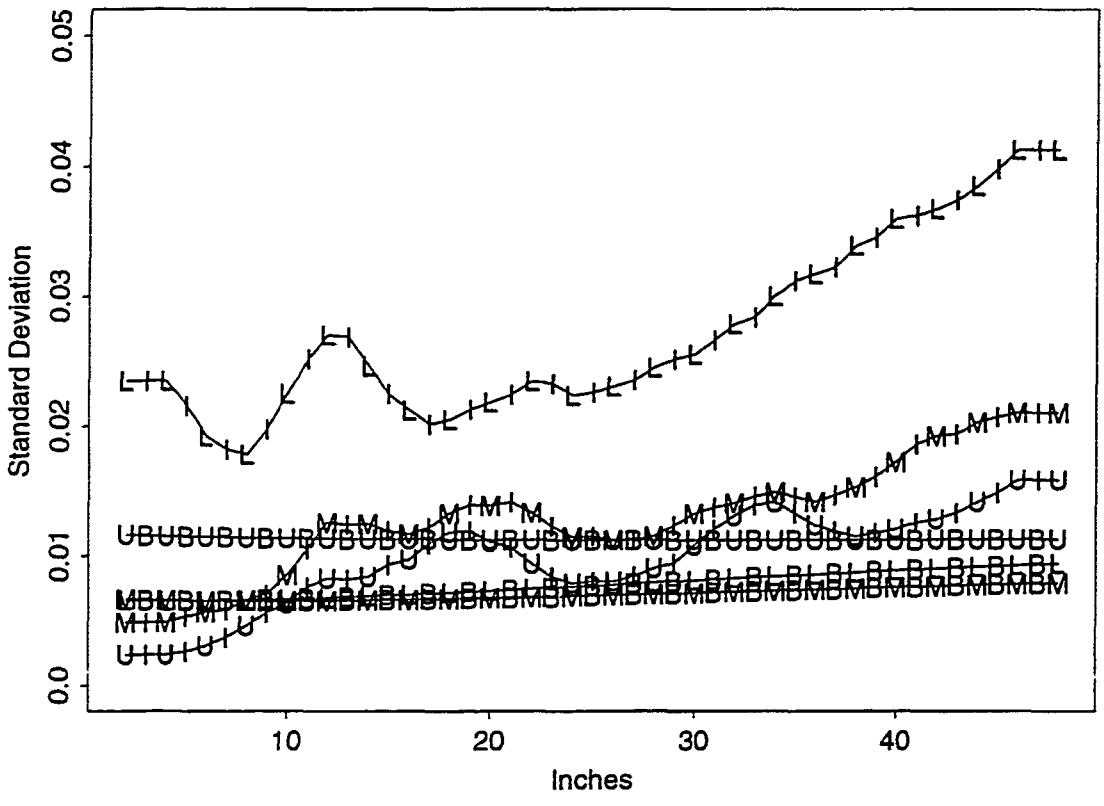

Figure 6.8 Comparison of estimated profiles of standard deviations of three quantiles: 'M' is median, 'U' is the upper quartile and 'L' is the lower quartile; 'B' is the estimated profile from the Bayesian approach and 'I' is the estimated profile from the imputation approach.

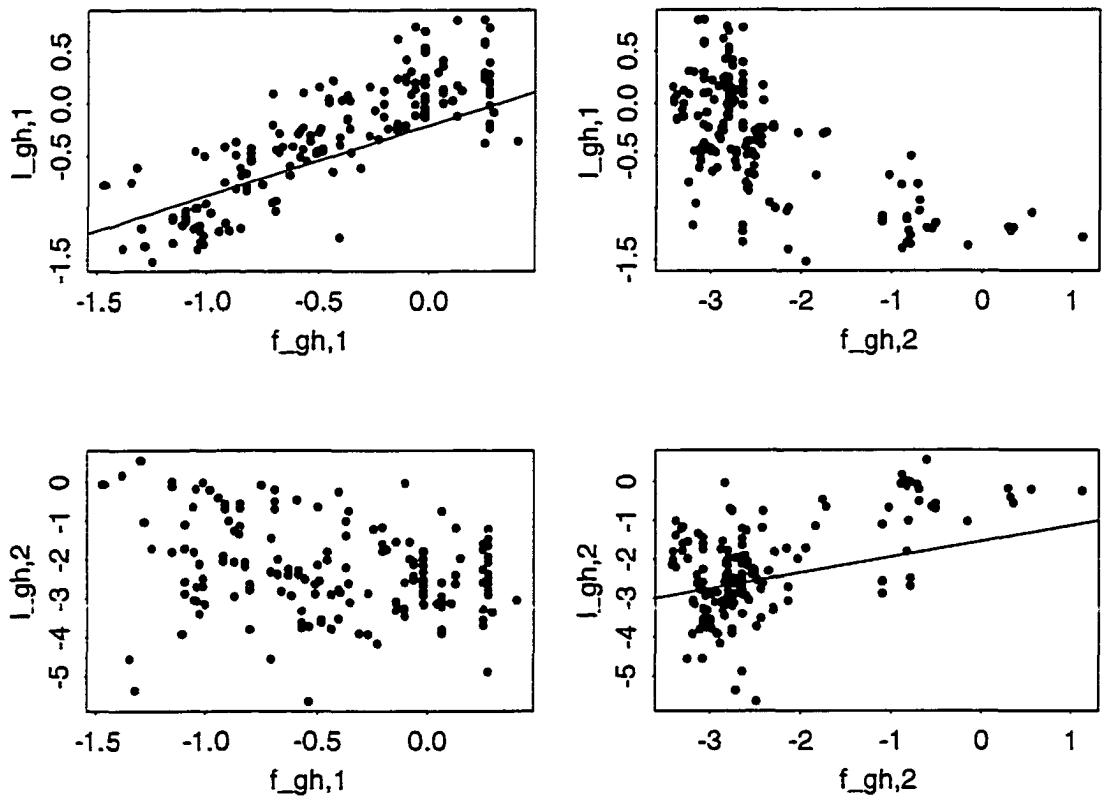

Figure 6.9 Scatter plot of components of f_{gh} versus those of l_{gh}.

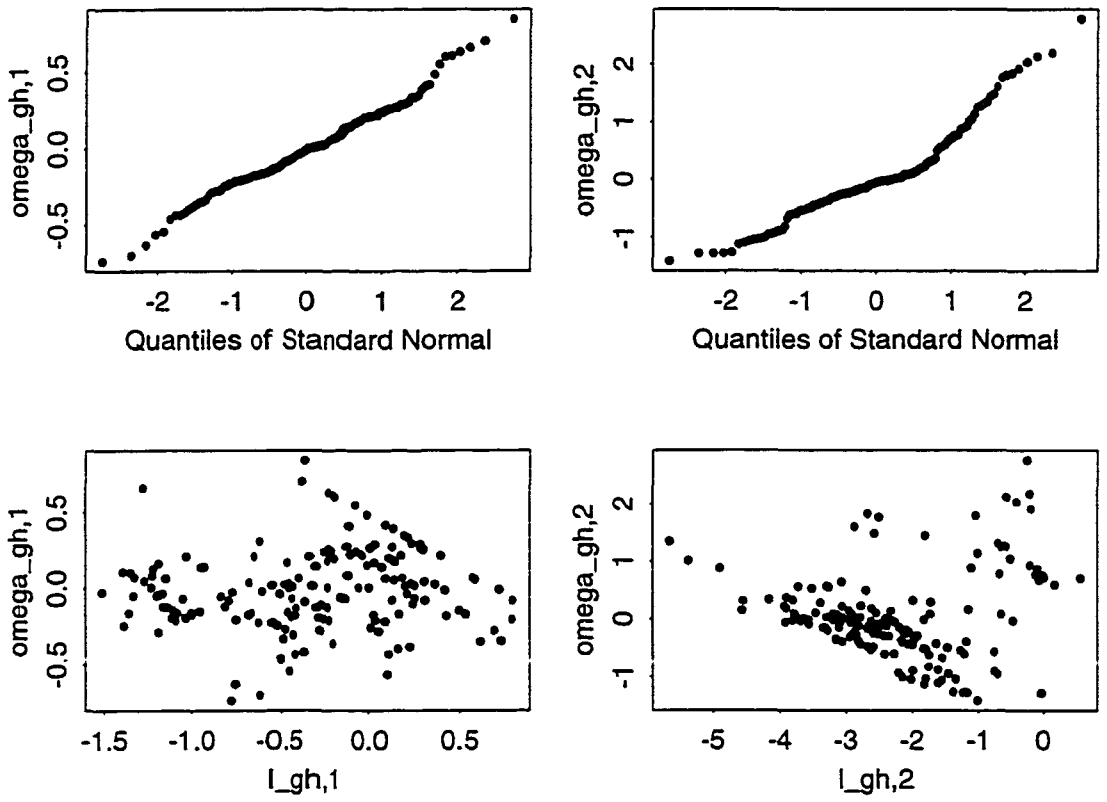

Figure 6.10 Residual plots for field measurement model. First row contains normal probability plots for each component of ω. Second row contains residual plots of each component of ω versus the corresponding component of f.

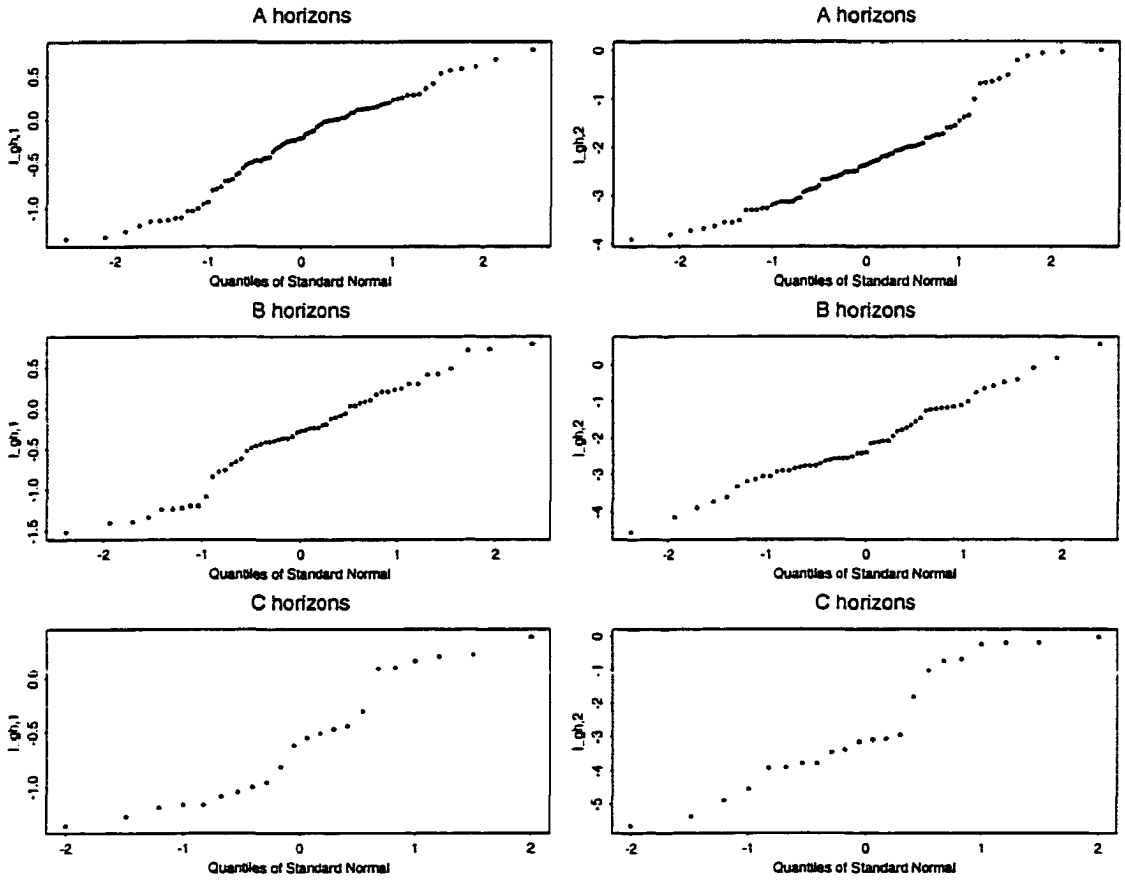

Figure 6.11 Normal probability plots of l_{gh}.

A simplifying assumption of the field measurement model is that ψ is the same for all sites $g \in \mathcal{D}$. To assess the validity of this assumption, we define a discrepancy measure

$$D_1 = \max_{g \in \mathcal{L}} \left(\sum_{h=1}^{H_g} \left[\boldsymbol{L}'_{gh} \boldsymbol{L}_{gh} \right]^{-1} \left[\boldsymbol{L}'_{gh} \boldsymbol{f}_{gh} \right] \right) - \min_{g \in \mathcal{L}} \left(\sum_{h=1}^{H_g} \left[\boldsymbol{L}'_{gh} \boldsymbol{L}_{gh} \right]^{-1} \left[\boldsymbol{L}'_{gh} \boldsymbol{f}_{gh} \right] \right),$$

where max and min are element-wise maximum and minimum. This measure is constructed to detect large differences in site by site estimates of the regression coefficients. The value of D_1 is calculated for the original data and for each replicate sample. The replicate samples are created by drawing new values of $\boldsymbol{f}_{gh}$ and $\boldsymbol{l}_{gh}$ for each observed horizon. Note that D_1 is a true statistic; it does not depend on parameters.

Figure 6.12 shows the estimated distribution of D_1 under the model. The solid vertical line in each plot is the value of D_1 for the original data. The dotted line in two of the plots indicates where the realized discrepancy would lie except for three sites. These sites seem to be extreme as measured by this discrepancy. Without these three sites, the replicate data appear to be reasonably similar to the original data, except for possibly ψ_{22}. The sample size here is not large enough to allow us to remove three "outliers", but it suggests a starting place for generalizing the model to fit the data better.

Another simplifying assumption of the model is that Σ_ω is constant. This assumption is investigated using a method similar to above. We define

$$D_2 = \max_{g \in \mathcal{L}} \left(H_g^{-1} \sum_{h=1}^{H_g} \left[\boldsymbol{f}_{gh} - \psi \boldsymbol{L}_{gh} \right] \left[\boldsymbol{f}_{gh} - \psi \boldsymbol{L}_{gh} \right]' \right) - \min_{g \in \mathcal{L}} \left(H_g^{-1} \sum_{h=1}^{H_g} \left[\boldsymbol{f}_{gh} - \psi \boldsymbol{L}_{gh} \right] \left[\boldsymbol{f}_{gh} - \psi \boldsymbol{L}_{gh} \right]' \right)$$

Note that D_2 is a function of the parameter ψ. Thus for each draw, $\theta^{(t)}$, the discrepancy is calculated for the original data and for the replicate data.

Figure 6.13 contains plots D_2 for both diagonal elements of Σ_ω. In the first plot, the replicate data resembles the original data. However, in the second plot, D_2 for the original data is generally higher than that for the replicate data. This may indicate that the model is oversimplified. Recall that calibration groups were developed in the

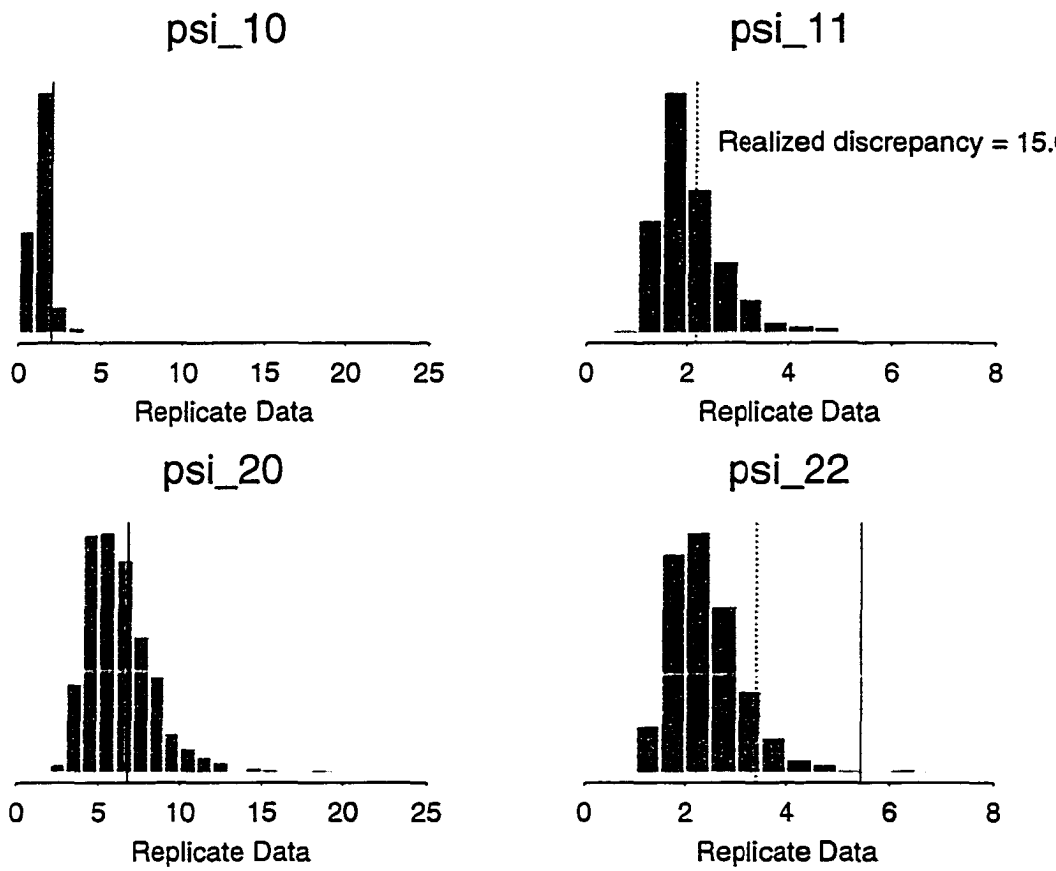

Figure 6.12 Estimated distribution of D_1.

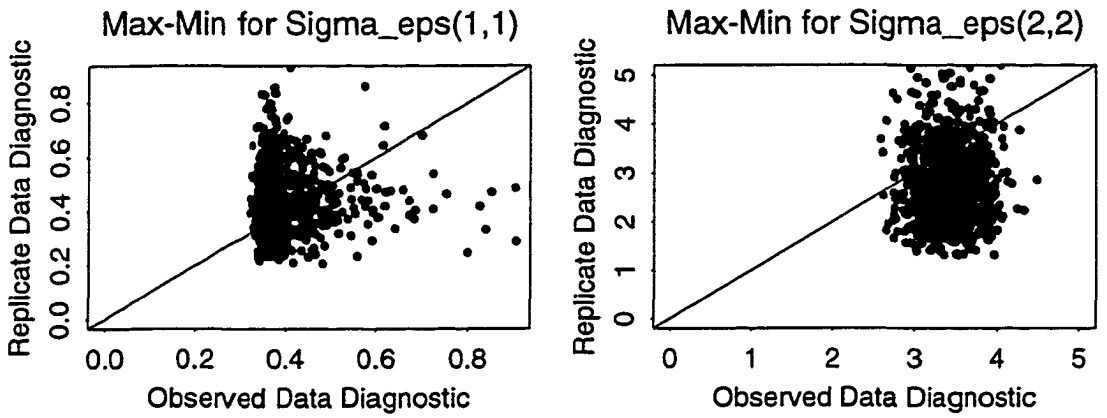

Figure 6.13 Estimated distribution of D_2.

imputation approach. However, we are using data from only one of those calibration groups. In general, the field and laboratory measurement models could be improved but seem to fit adequately.

6.8.2 Horizon profile model

Figure 6.14 shows empirical transition probabilities for all inches for the nine free parameters of $\mathbf{\Delta}$. In these plots, there is some evidence of inhomogeneity of the Markov chain across depth. The empirical transition probability at each inch can be used as a discrepancy measure for posterior predictive assessment. Define

$$D_3(i,j,k) = \frac{\sum_g \mathbb{I}\left(T_{g,i-1}=j\right)\mathbb{I}\left(T_{gi}=k\right)}{\sum_g \mathbb{I}\left(T_{g,i-1}=j\right)}. \tag{6.11}$$

Figure 6.15 shows values of D_3 for selected inches. The approximate posterior distribution of D_3 under the model is displayed as a box plot for selected inches. Next to each box plot is a single observation corresponding to the empirical transition probability in the original data. In many cases, D_3 for the original data falls in the tail of the distribution from the replicated data.

As a simple extension of the model, we allow the matrix $\mathbf{\Delta}$ to be piecewise constant for intervals of six inches. Figure 6.16 shows posterior means of the nine free parameters of $\mathbf{\Delta}$ under the piecewise constant model. Note that in bins with no data, the prior

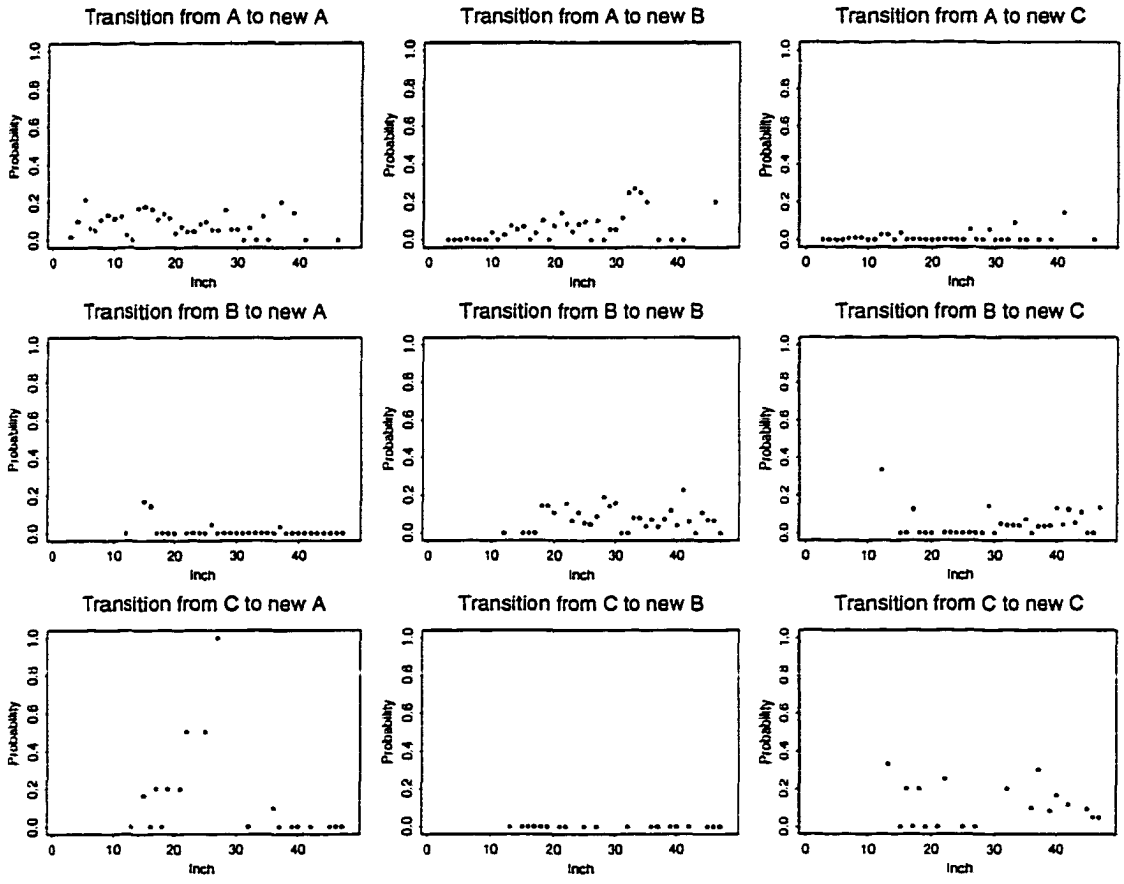

Figure 6.14 Empirical transition probabilities for data for all inches.

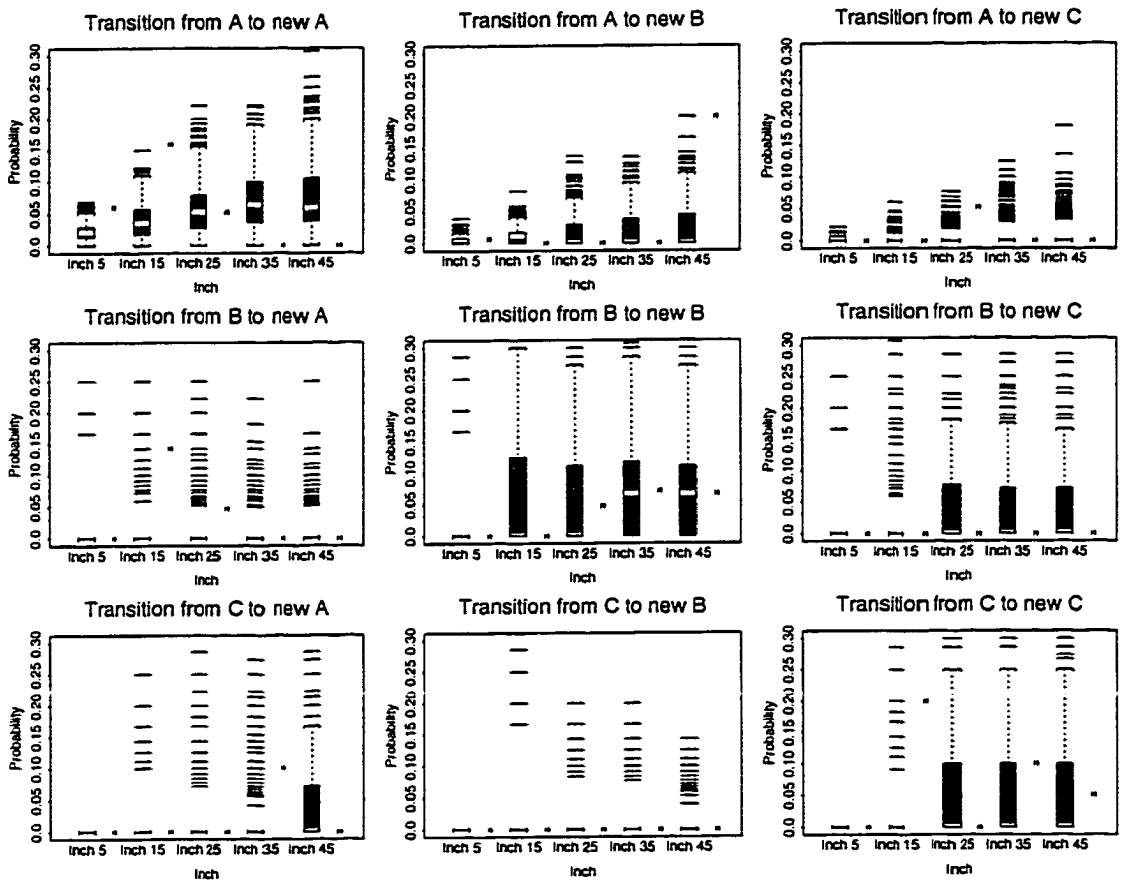

Figure 6.15 Discrepancy measures for transition probabilities for replicate and original data for selected inches.

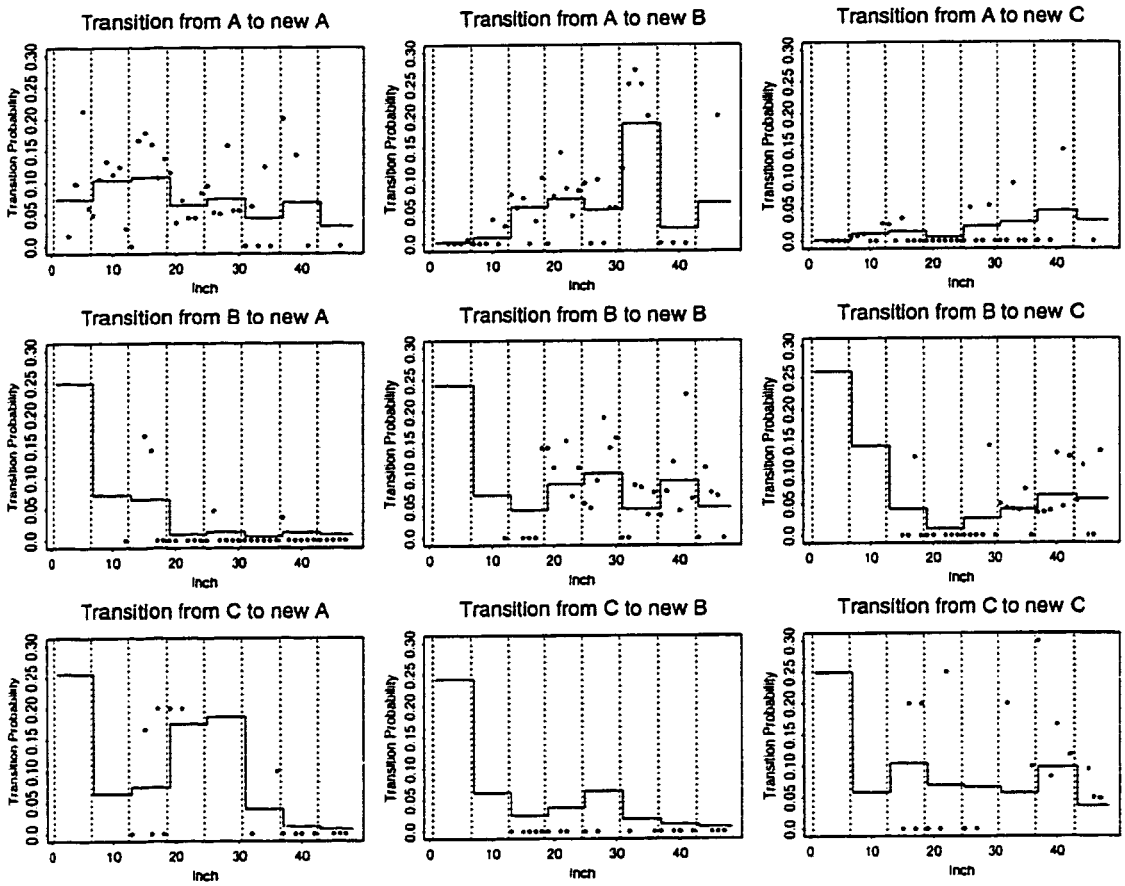

Figure 6.16 Posterior means of piecewise constant transition probabilities (solid line) with discrepancy measures for original sample (dots). Vertical dotted lines represent intervals on which Δ is constant.

mean (0.25 for each transition probability) is the posterior mean. For example, in the first bin for the elements of δ_B, the posterior means are 0.25.

In general, the piecewise constant posterior means seem to fit the empirical transition probabilities better. Using the same discrepancy measure for data replicated under the piecewise constant model, Figure 6.17 shows that the fit of the horizon profile is substantially improved in many cases, but there remains evidence of lack of fit. In Section 6.9, we suggest an extension which allows the matrix Δ to change smoothly across inches. This extension may further improve the fit of the horizon profile model.

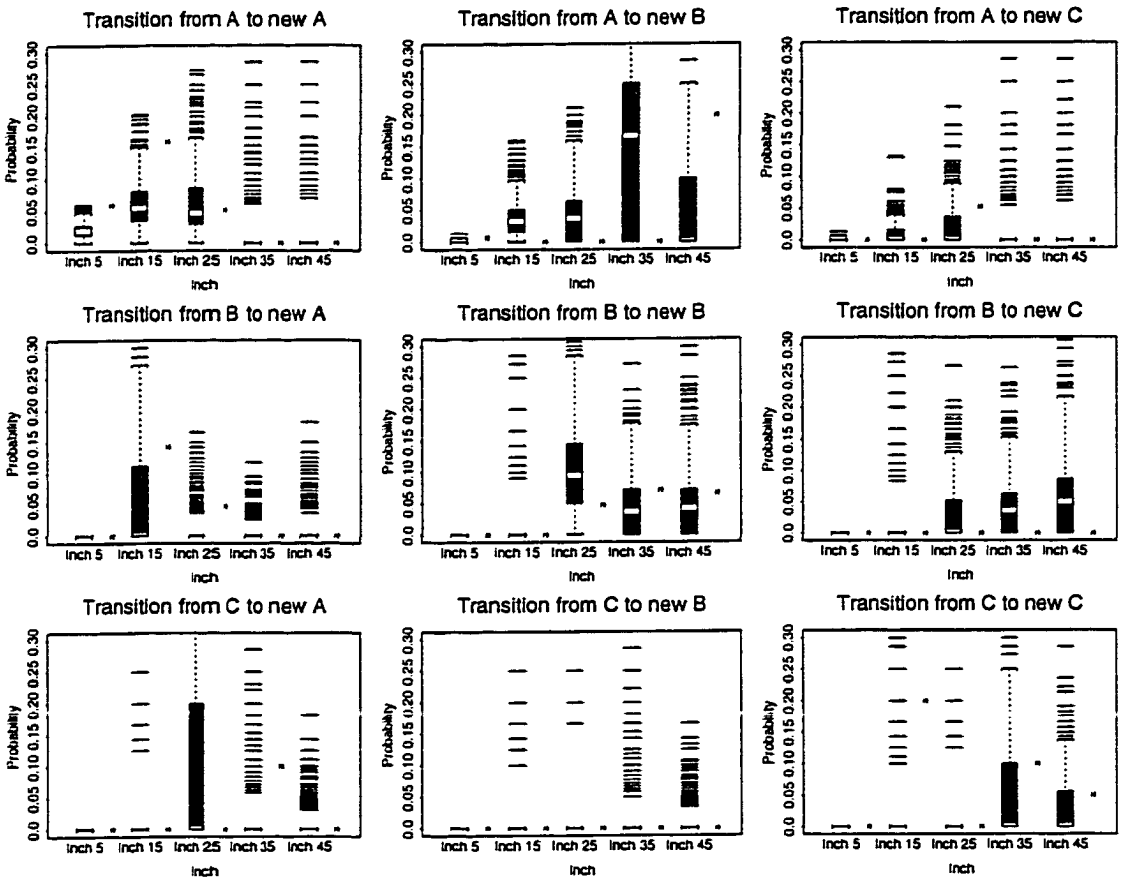

Figure 6.17 Discrepancy measures for transition probabilities for replicate and original data for selected inches under the piecewise constant model.

6.9 Improvement of the hierarchical model

6.9.1 Use of a Box-Cox transformation

To address possible inadequacies of the field and laboratory measurement model, we consider a larger class of transformations. Iyengar and Dey (1998) consider this class of transformations for compositional data in a Bayesian analysis. Let

$$l_{gh}(\tau) = \begin{cases} \mathbb{L}(c_1^{(l)}, c_2^{(l)}, c_3^{(l)}) & \text{if } \tau = 0 \\ \left(\left(\frac{c_1^{(l)}}{c_3^{(l)}} \right)^{\tau}, \left(\frac{c_2^{(l)}}{c_3^{(l)}} \right)^{\tau} \right) & \text{otherwise,} \end{cases} \tag{6.12}$$

Define $f_{gh}(\tau)$ similarly. Then τ becomes another parameter of the model. We might search for a transformation such that $l_{gh}(\tau)$ is more normal than under the log-ratio transformation. In choosing from this larger class of transformations, we may be able to improve the fit of the data model. Recall that an adjustment vector was used to ensure that all components of texture are positive. The adjustment vector could also be viewed as a parameter of the model.

6.9.2 Heterogeneous Markov chain model

The greatest lack of fit was found in the model for the transition probabilities. The extended model allows the transition probabilities of the Markov Chain to vary with depth. Extend the previous notation such that $\mathbb{P}(T_{gi} = k \,|\, T_{g,i-1} = j) = \delta(j,k,i)$. Define $\boldsymbol{\Delta}_i$ to be the transition matrix containing these probabilities. This matrix will have the same form as (6.3).

In order to model the elements of $\boldsymbol{\Delta}_i$ as functions of depth, we use a log-ratio transformation similar to the one presented in (3.1). The Dirichlet distribution used for the prior distributions for $\boldsymbol{\delta}_A$, $\boldsymbol{\delta}_B$ and $\boldsymbol{\delta}_C$ in the original model does not accommodate covariates. For $j = 1,3,5$ and $i = 2, \ldots, 48$, define

$$\boldsymbol{\lambda}_{ji} = \left(\log \frac{\delta(j,2,i)}{\delta(j,j,i)}, \log \frac{\delta(j,4,i)}{\delta(j,j,i)}, \log \frac{\delta(j,6,i)}{\delta(j,j,i)} \right).$$

Note that $\boldsymbol{\lambda}_{1i}$ is a transformation of the (extension of the) elements of $\boldsymbol{\delta}_A$. Similarly, $\boldsymbol{\lambda}_{3i}$ corresponds to $\boldsymbol{\delta}_B$ and $\boldsymbol{\lambda}_{5i}$ corresponds to $\boldsymbol{\delta}_C$. Then the back-transformation is defined

by

$$\begin{aligned}
\delta(j,2,i) &= \frac{\exp(\lambda_{ji,1})}{1+\exp(\lambda_{ji,1})+\exp(\lambda_{ji,2})+\exp(\lambda_{ji,3})}, \\
\delta(j,4,i) &= \frac{\exp(\lambda_{ji,2})}{1+\exp(\lambda_{ji,1})+\exp(\lambda_{ji,2})+\exp(\lambda_{ji,3})}, \\
\delta(j,6,i) &= \frac{\exp(\lambda_{ji,3})}{1+\exp(\lambda_{ji,1})+\exp(\lambda_{ji,2})+\exp(\lambda_{ji,3})}, \\
\delta(j,j,i) &= \frac{1}{1+\exp(\lambda_{ji,1})+\exp(\lambda_{ji,2})+\exp(\lambda_{ji,3})}.
\end{aligned} \tag{6.13}$$

We model $\boldsymbol{\lambda}_{ji}$ as a quadratic function of depth. Let

$$\boldsymbol{\lambda}_{ji} = \begin{pmatrix} 1 & i & i^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & i & i^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & i & i^2 \end{pmatrix} \boldsymbol{\phi}_j = \boldsymbol{Z}_i \boldsymbol{\phi}_j. \tag{6.14}$$

A normal prior will be used for $\boldsymbol{\phi}_j$ with the notation

$$\boldsymbol{\phi}_j \sim \mathbb{N}(\boldsymbol{D}_j, \boldsymbol{V}_D).$$

In order to derive the conditional posterior of $\boldsymbol{\phi}_j$, we consider the distribution of the Markov Chain as a function of the newly defined parameters. The augmented Markov Chain data are considered. The previous indicator notation is modified so that $\mathcal{T}_{jgi} = \mathrm{I}\{T_{gi} = j\}$. Then

$$p(\{\boldsymbol{T}_g\}_{g\in\mathcal{D}} \mid \boldsymbol{\Delta}_i) = \prod_{g\in\mathcal{D}} \prod_{i=2}^{\min\{I_g+1,48\}} \prod_{j=1}^{6} \prod_{k=1}^{6} \delta(j,k,i)^{\mathcal{T}_{kgi}\mathcal{T}_{jg,i-1}}.$$

Due to the structure of the matrix $\boldsymbol{\Delta}$, the products over j and k can be reduced to products over the sets $j \in \{1,3,5\}$ and $k \in \{j,2,4,6\}$. Further algebra leads to

$$\begin{aligned}
&\prod_{g\in\mathcal{D}} \prod_{i=2}^{\min(I_g+1,48)} \prod_{j\in\{1,3,5\}} \prod_{k\in\{j,2,4,6\}} \left(\frac{\delta(j,k,i)}{\delta(j,j,i)}\right)^{\mathrm{I}(T_{gi}=k,T_{g,i-1}=j)} \delta(j,j,i)^{\mathrm{I}(T_{gi}=k,T_{g,i-1}=j)} \\
&= \prod_{g\in\mathcal{D}} \prod_{i=2}^{\min(I_g+1,48)} \prod_{j\in\{1,3,5\}} \delta(j,j,i)^{\mathrm{I}(T_{g,i-1}=j)} \prod_{k\in\{j,2,4,6\}} \exp\left\{\mathbb{I}(T_{gi}=k, T_{g,i-1}=j) \log\frac{\delta(j,k,i)}{\delta(j,j,i)}\right\} \\
&= \prod_{g\in\mathcal{D}} \prod_{i=2}^{\min(I_g+1,48)} \prod_{j\in\{1,3,5\}} \delta(j,j,i)^{\mathrm{I}(T_{g,i-1}=j)} \exp\left\{\mathbb{I}(T_{g,i-1}=j)\,\lambda_{ji,(T_{gi}/2)}\right\}.
\end{aligned}$$

Using equation (6.14), we have

$$\begin{aligned} p(\boldsymbol{\phi}_j \,|\, \cdot) \;&\propto\; \prod_{g\in\mathcal{D}} \prod_{i=2}^{\min(I_g+1,48)} \delta(j,j,i)^{\mathbb{I}(T_{g,i-1}=j)} \exp\left\{\mathbb{I}\left(T_{g,i-1}=j\right)\lambda_{ji,(T_{gi}/2)}\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}(\boldsymbol{\phi}_j-\boldsymbol{D}_j)'\boldsymbol{V}_D(\boldsymbol{\phi}_j-\boldsymbol{D}_j)\right\} \\ &\propto\; C_j \exp\left\{\left[\left(\sum_i \boldsymbol{n}'_{ij}\boldsymbol{Z}_i\right)\boldsymbol{V}_D+\boldsymbol{D}'_j\right]\boldsymbol{V}_D^{-1}\boldsymbol{\phi}_j-\frac{1}{2}\boldsymbol{\phi}'_j\boldsymbol{V}_D^{-1}\boldsymbol{\phi}_j\right\}, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{n}'_{ij} \;&=\; \sum_{g\in\mathcal{D}} \mathbb{I}\left(T_{g,i-1}=j\right)\left\langle \mathbb{I}\left(T_{gi}=2\right), \mathbb{I}\left(T_{gi}=4\right), \mathbb{I}\left(T_{gi}=6\right)\right\rangle \\ \text{and } C_j \;&=\; \prod_{g,i} \delta(j,j,i)^{\mathbb{I}(T_{g,i-1}=j)}. \end{aligned}$$

This is not the kernel of a standard distribution. Thus, we cannot sample from it directly. However, we may be able to use rejection sampling or a Metropolis step to simulate draws from this distribution.

Note that $C_j \leq 1$. The other factor of the conditional posterior is the kernel of the density of a $\mathbb{N}\left(\boldsymbol{V}_D\left(\sum_i \boldsymbol{Z}'_i\boldsymbol{n}_{ij}\right)+\boldsymbol{D}_j, \boldsymbol{V}_D\right)$ random variable. Denote this density by $\varphi(\boldsymbol{\phi}_j)$. We have that the importance ratio $p(\boldsymbol{\phi}_j \,|\, \cdot)/\varphi(\boldsymbol{\phi}_j) = C_j$ is bounded by 1. In order to carry out rejection sampling, we will draw a candidate $\boldsymbol{\phi}_j$ from p'. This candidate will be accepted with probability C_j. We will use the fact that

$$\begin{aligned} C_j \;&=\; \prod_{g\in\mathcal{D}} \prod_{i=2}^{\min(I_g+1,48)} \left[\exp\left(\frac{1}{\langle 1,1,1\rangle \boldsymbol{Z}_i\boldsymbol{\phi}_j}\right)\right]^{\mathbb{I}(T_{g,i-1}=j)} \\ &=\; \exp\left(\sum_{g\in\mathcal{D}} \sum_{i=2}^{\min(I_g+1,48)} \frac{\mathbb{I}\left(T_{g,i-1}=j\right)}{1+\langle 1,1,1\rangle \boldsymbol{Z}_i\boldsymbol{\phi}_j}\right). \end{aligned}$$

6.10 Conclusion

In this chapter, we presented a hierarchical model for analyzing the soil texture data. The distributional assumptions in the field and laboratory measurements model and the horizon profile model allow us to derive the posterior distribution of the quantile profiles.

We are able to use Gibbs sampling to simulate from this distribution. Inferences about the quantile profiles can then be made using the draws from the Gibbs sampler. Having the estimated posterior distribution provides more information about the behavior of this quantile estimator than the jackknife variance estimates provide for the CD-based quantile estimator.

This model can be used for analyses of many other variables in the pilot project. Other horizon-based measurements can be modeled as functions of the horizon profile as in the laboratory measurement model. Some variables of interest can be defined as functions of the transition probabilities of the horizon profile model, e.g., depth of the surface horizon and thickness of the B horizon.

7 COMPARISON OF METHODOLOGIES

7.1 Introduction

The two analysis approaches presented in Chapters 5 and 6 both produce estimated quantile profiles. In this chapter, we compare these approaches on the basis of modeling assumptions, computational aspects and the output produced by each procedure. A simulation study is presented in which data generated under the hierarchical model is analyzed using the imputation approach. Because the complete model in the imputation approach is not explicit, it is difficult to simulate data for Bayesian analysis for comparison.

7.2 Modeling assumptions

In the imputation approach presented in Chapter 5, the assumptions of the calibration and imputation models are fairly mild. In the calibration step of the procedure, we assume a linear relationship between transformed laboratory measurements, $\boldsymbol{l}_{gh}$, and and transformed field measurements, $\boldsymbol{f}_{gh}$, within each calibration group. Since errors in these predictions are ignored in the imputation step, the calibration model must fit well in order to obtain reasonable quantile estimates. As we will show in the simulations, even a good calibration model can cause bias in the final quantile estimator.

In the imputation step, we assume that there is a linear relationship between the calibrated laboratory value for inch i, $\hat{\boldsymbol{l}}_{ghi}$, and the calibrated laboratory value for the first inch, $\hat{\boldsymbol{l}}_{gh1}$, for $i = 2, \ldots, 48$. We also assume that, for each inch, residuals from this regression are homoskedastic within imputation classes. The imputation classes were

developed to attempt to make these assumptions reasonable.

After estimating the marginal distribution functions for each component of texture, we invert the functions to obtain marginal quantile estimates. The marginal quantile estimates are smoothed across inches using a simple moving average of five inches. In doing the smoothing, we assume that the population quantiles are changing smoothly across inches. At each site, the values of texture may change abruptly, indicated by a boundary between horizons. However, averaging across sites, we expect that the distribution of texture changes smoothly across inches.

The imputation approach can be characterized as a semi-parametric method. Explicit distributional assumptions are not required. On the other hand, the modeling assumptions of the hierarchical model are strong and specific. Specific modeling assumptions provide a formal basis for diagnostics and an explicit simulation platform for evaluating estimation procedures. The lack of explicit assumptions in the imputation approach leads to a lack of falsifiability of the model.

In the hierarchical model, we assume a linear relationship between $\boldsymbol{l}_{gh}$ and $\boldsymbol{f}_{gh}$, as in the imputation approach. However, we further assume that the residuals from a regression of $\boldsymbol{f}_{gh}$ on $\boldsymbol{l}_{gh}$ follow a normal distribution. In the next level of the model, $\{\boldsymbol{l}_{gh}\}$ are assumed to be distributed normally with a mean and variance depending on the master horizon designation. If these assumptions are approximately satisfied, then meaningful inferences can be made from the posterior distribution.

The hierarchical model also assumes a distributional structure for horizon profiles. As evidenced by the data collection protocol, the horizons found at a particular site greatly influence the profile of most soil characteristics. The distribution of the horizon profiles will be applicable in modeling many of the other variables of interest in the MLRA 107 project, in particular, variables directly related to the horizon profile such as, depth to first B horizon, depth of the A horizon, etc. This is a strength of the Bayesian approach, because it will allow a unified analysis approach to many variables, as well as for other horizon-based measurements.

7.3 Computational aspects

In the imputation approach, the computations required are fairly simple. A collection of linear regression models must be fit. Predictions and residuals from these models are needed. The calculation of the weighted empirical distribution function involves only a weighted sum of indicators. This step function is inverted to obtain marginal quantile estimates. The calculation of smoothed quantile estimates is also very simple.

The Bayesian analysis involves more complex computation. Gibbs sampling can be used to sample from the posterior distribution of the parameters given the data. In each iteration of the Gibbs sampling, we must sample from the conditional posteriors for each subvector of the parameter vector. While this is more complex than the computations required for the imputation approach, this particular hierarchical model allows a relatively simple way of sampling from the posterior distribution. The extension to a heterogeneous Markov chain model for the horizon profile is slightly more complex, but still computationally feasible.

7.4 Output

There is a trade-off for mild assumptions and computational simplicity. This is reflected in the richness of the output from each of the approaches. To assess the quality of the estimated quantile profiles, jackknife variance estimates are computed.

Under the hierarchical model, we can obtain the posterior distribution for the quantile profiles given the data, not just variance estimates. The Bayesian analysis also yields posterior distributions for all of the parameters in the model and any transformation of these parameters.

In the data analyzed in Section 6.7, the normality assumptions for the log-ratio transformed field and laboratory measurements appeared reasonable for the data. If this had not been the case, the richness of the output from the Bayesian analysis is incentive to find a transformation for which the normality assumptions do appear to hold. One way to search for such a transformation is by including a parameter for a

Box-Cox type transformation for laboratory and field measurements as mentioned in Section 6.9 (Iyengar and Dey, 1998).

7.5 Simulation

We can simulate data from the hierarchical model to evaluate the performance of the imputation approach for data with similar structure to the soil texture data for which the true distribution is known. Some modifications of the imputation approach were considered to reduce bias and maximum likelihood estimates were calculated. Clay quantile estimates are obtained at the .01, .05, .10, .25, .50, .75, .90, .95 and .99 quantiles to demonstrate the properties of $\tilde{Q}$.

The imputation approach was applied to 100 replications of simulated data. The structure of the simulated data is similar to that of the soil texture data. The sample sizes are $|\mathcal{S}| = 60$ (surface horizon sites), $|\mathcal{F}| = 10$ (full profile sites) and $|\mathcal{L}| = 10$ (laboratory sites). Data were generated under a set of parameters similar to the posterior means from the analysis of Chapter 6. However, the variance of the site-specific random effect was set to zero. Because the imputation approach does not explicitly account for a site-specific random effect, it was expected to perform better for data without a random effect.

Figure 7.1 shows the height of the density at several quantiles for selected inches. Note that at inch 2, the density is severely skewed, but becomes more symmetric toward the bottom of the profile. However, the density has a long right tail throughout the entire profile.

The estimates ($\tilde{Q}$ for shorthand notation) are obtained from the imputation approach as presented in Chapter 5. No calibration groups or imputation classes were considered since the data were not generated with this structure. Maximum likelihood estimates (MLEs) can be calculated. Since the likelihood is correct for the simulated data, we expect the MLEs to perform very well.

Relative bias is a measure commonly used to assess the performance of an estimator.

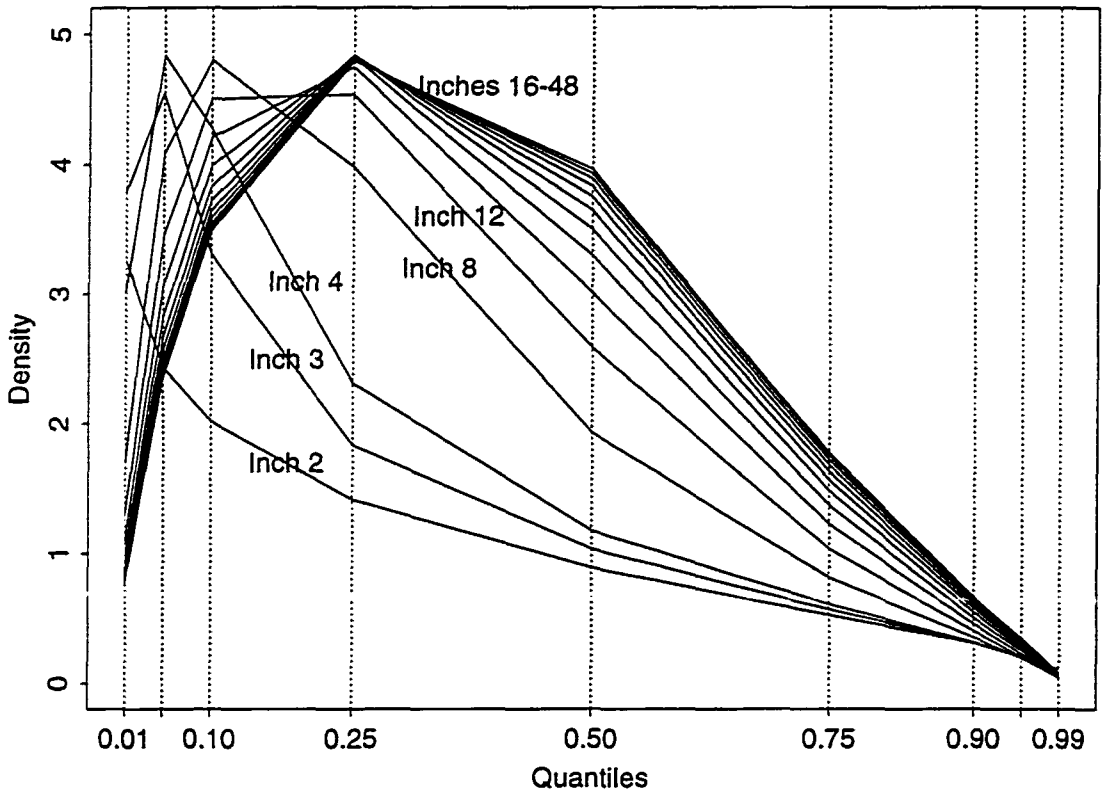

Figure 7.1 Probability density functions of clay content at selected inches.

The relative bias of an estimator $\tilde{Q}$ as an estimator of a quantity, Q is

$$\frac{\tilde{Q}-Q}{Q}.$$

Figure 7.2 contains profiles of the relative bias of $\tilde{Q}$ (solid line) and the relative bias of the MLE (dashed line) for the nine quantiles. The horizontal (dotted) line at each quantile represents 0% relative bias for an estimate of that quantile. The area above the line is positive bias; the area below is negative bias. For example, the $\tilde{Q}(.01)$ is negatively biased for the first two inches and then becomes positively biased for inch 3 and onward. The plot is scaled so that the 0% reference line for one quantile also represents the 100% reference line for the previous quantile and the -100% reference line for the following quantile.

The bias of the MLE falls nearly on top of the 0% reference line for each quantile, indicating that the MLE has essentially no bias. However, $\tilde{Q}$ seems to have a large bias which is worse in the left tail of the distribution. The sign of the bias indicates that, for most of the profile, the estimated distribution is much more peaked than the true distribution. In the first two inches of the profile, the bias for all estimated quantiles is negative. This may be due to the fact that the density is more severely skewed in the first few inches.

Because of the large bias observed in the simulation, we investigated bias that may result from the imputation step of the estimation procedure. In the imputation step, imputing multiple values should produce data with approximately the same amount of variability as the true population. However, the variance of the fitted residuals underestimates the variance of the true population of residuals by a factor of $n^{-1}(n-2)$. For large samples, this factor is near 1.0, but for the sample sizes in the simulated data, it is 0.8 near the bottom of the profile. This suggests that an inflation factor might be used to increase the variability in the fitted residuals to the right amount. This approach is used in the context of bootstrapping regression models (e.g., Stine, 1985).

In the imputation step, we use $\sqrt{n(n-2)^{-1}}\hat{E}$ as the residual instead of $\hat{E}$. In the first few inches of the profile, this adjustment is quite small since the sample size is near 80. There is a stronger effect near the bottom of the profile where the sample size is 10.

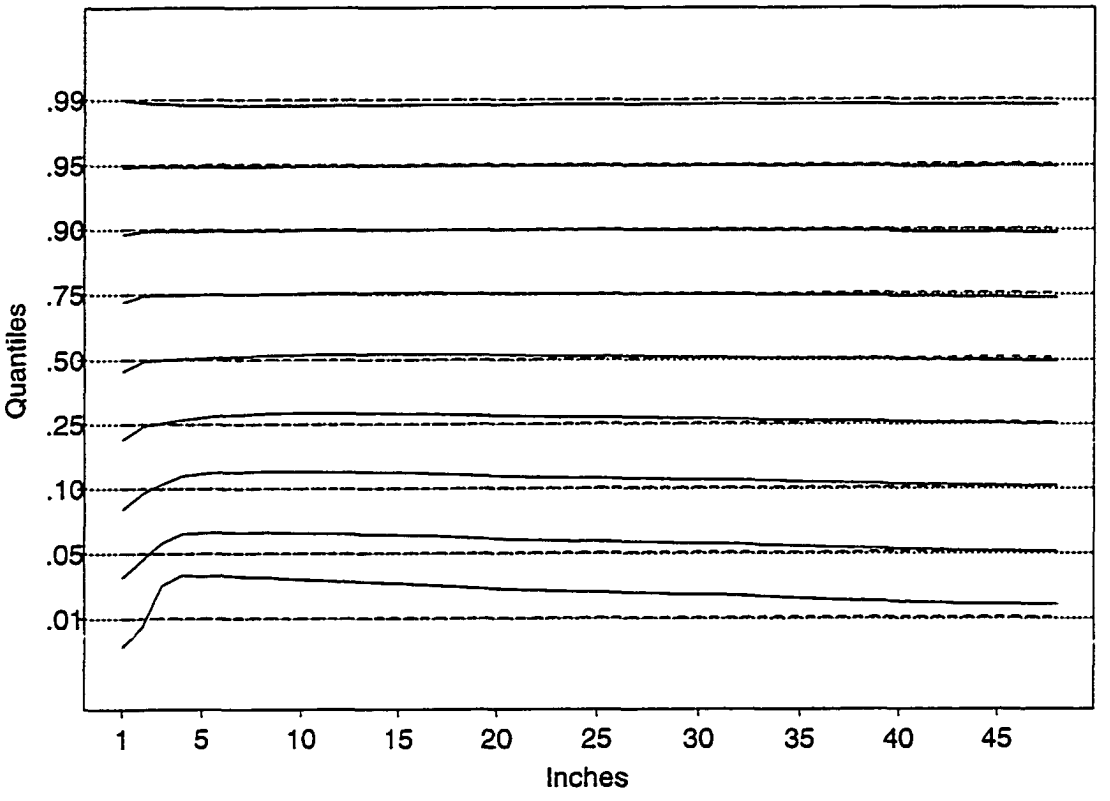

Figure 7.2 Relative bias of $\tilde{Q}$ and MLE. Solid lines are relative bias of $\tilde{Q}$; dashed lines are relative bias of MLEs; dotted lines are reference lines as described on page 155.

Figure 7.3 shows the relative bias of the quantile estimates using the inflated residuals (solid line) versus that of $\tilde{Q}$ (dashed line). This plot can be interpreted as described on page 155. In general, the absolute relative bias is reduced, but not considerably. Thus, we investigated the calibration step for its effect on the bias.

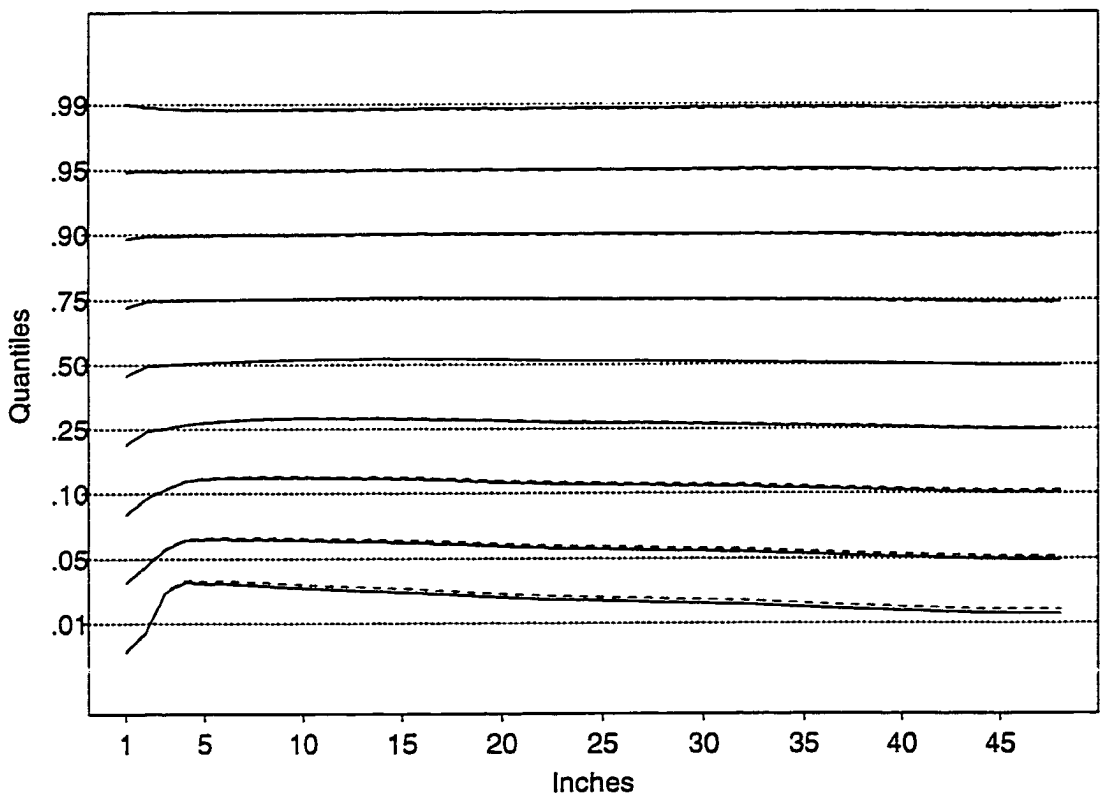

Figure 7.3 Relative bias of the quantile estimates using an inflation factor. Solid lines are relative bias for estimates using inflation factor; dashed lines are relative bias of $\tilde{Q}$; dotted lines are reference lines as described on page 155.

In the calibration step, we use a naive imputation procedure. Error in the predictions from the calibrated models is ignored. To see the effect of calibration on the estimates we allow laboratory values to be known for every observed horizon. This essentially reduces the structure of the data to a two-phase structure and removes the need for the

calibration step in the estimation procedure. Figure 7.4 shows the relative bias of the quantile estimator without calibration (solid line) versus that of $\tilde{Q}$ (dashed line). The relative bias is greatly improved by removing the calibration step. In most cases, the absolute relative bias is reduced by 50%.

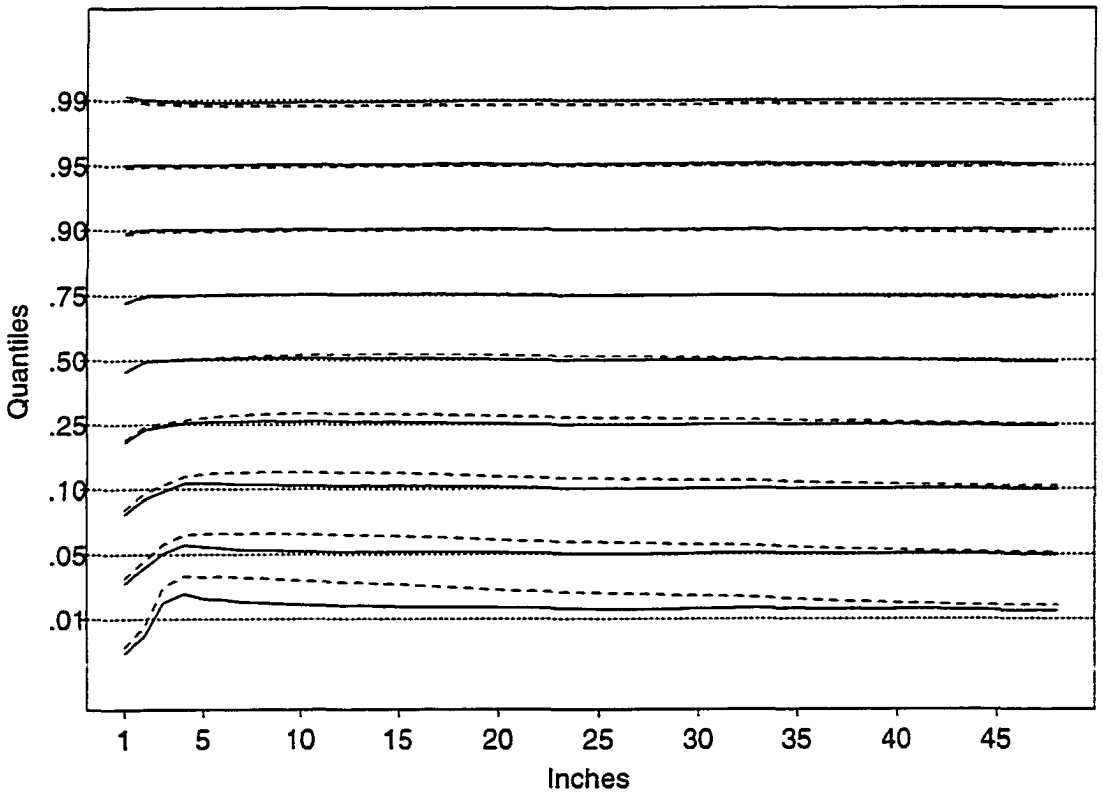

Figure 7.4 Relative bias of the quantile estimates without calibration. Solid lines are relative bias for estimates without calibration; dashed lines are relative bias of $\tilde{Q}$; dotted lines are reference lines as described on page 155.

While it is comforting to know that the calibration step is a large source of the bias we saw in the initial estimates, removing the calibration step is not an option for a real three-phase data set, such as the soils texture data. This type of bias is exactly what the Chambers and Dunstan imputation method aims to prevent. However, it is impractical

to do this type of imputation at both the calibration and imputation steps, because the resulting imputed data set would be huge.

To attempt to reduce the bias in a realistic way, we investigated a random calibration method. As before, we fit the calibration model and produce a prediction for each observed horizon. For each prediction, a fitted residual from the calibration model is randomly selected and added to the prediction. This randomly calibrated value can be thought of as a random sample from the imputed values we could have produced using the Chambers and Dunstan imputation procedure. Figure 7.5 shows the relative bias for the quantile estimator using random calibration (solid line) versus that of $\tilde{Q}$ (dashed line). The variance of the estimates does not change significantly using random calibration. Random imputation appears to be a practical solution which can greatly reduce the bias.

Returning to the original estimates, we compare the variance of $\tilde{Q}$ with that of the MLEs. Figure 7.6 contains profiles of the ratio of the variance of the MLE to the variance of the quantile estimator for the nine quantiles. The dotted lines represent a ratio of zero (variance of the MLE is zero) and one (variance of the MLE is equal to that of the quantile estimator). The plot shows that the ratio of the variances is almost always less than one, which indicates that the variance of the MLE is, in general, lower than that of the quantile estimator. This is expected, since the MLE is a parametric estimate based on the true model from which the data are generated. Throughout most of the profile, the ratio is less than 0.3 indicating that the variance of $\tilde{Q}$ is usually more than three times larger than that of the MLE.

Another benchmark for the variance of $\tilde{Q}$ is the asymptotic variance of the appropriate order statistic. This variance was presented in Chapter 4. Figure 7.1 shows the height of the density of clay content for each of the nine quantiles at selected inches. For a given quantile, an increase in the height of the density results in a decrease in the asymptotic variance of the order statistic.

A sample quantile, $Q_n(p)$ is a non-parametric estimator of $Q(p)$. If the assumptions of the imputation approach are true or at least reasonable, $\tilde{Q}$ should perform better

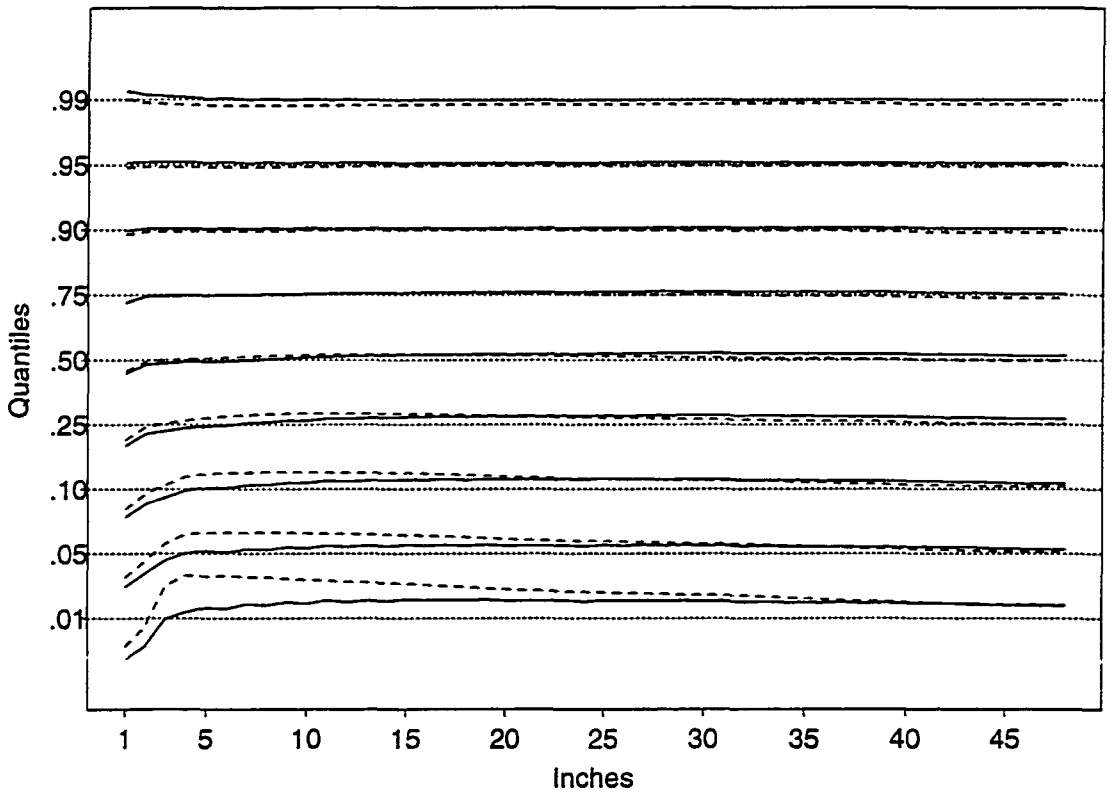

Figure 7.5 Relative bias of the quantile estimates using random calibration. Solid lines are relative bias for estimates using random calibration; dashed lines are relative bias of $\bar{Q}$; dotted lines are reference lines as described on page 155.

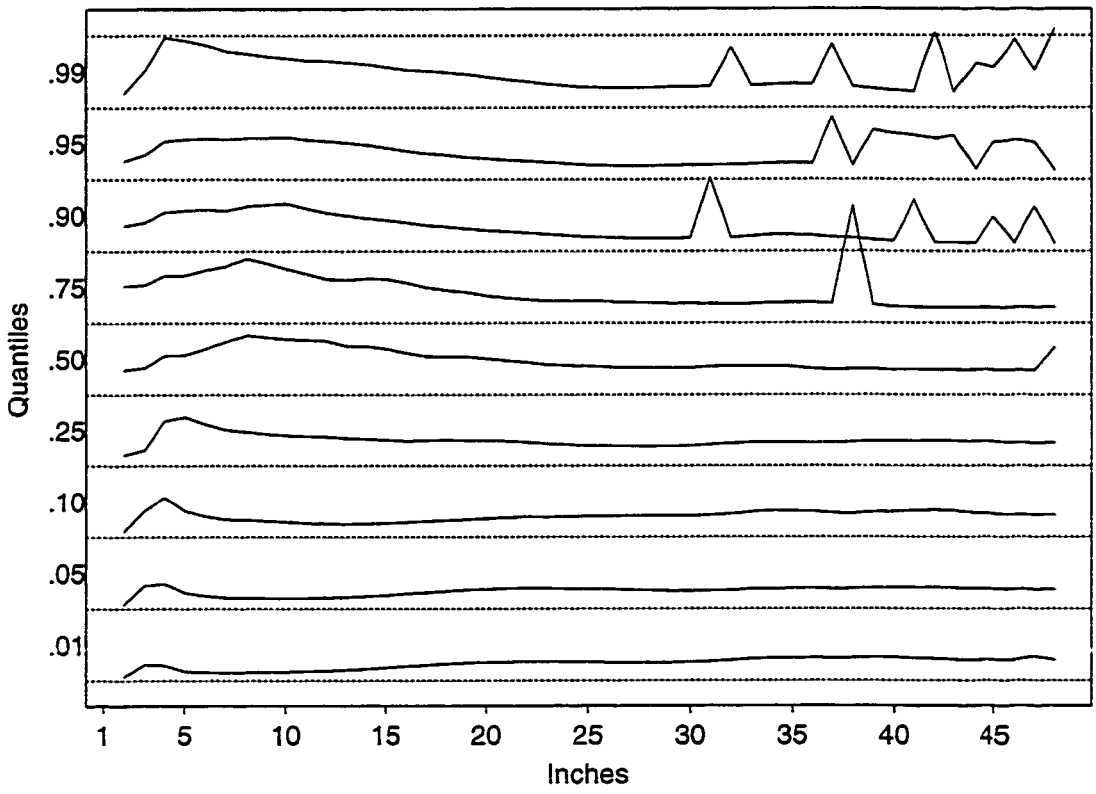

Figure 7.6 Ratio of variance of MLE to variance of $\tilde{Q}$. The bottom horizontal (dotted) line represents a ratio of 0 for $Q(.01)$. The next horizontal line represents a ratio of 1 for $Q(.01)$ and a ratio of 0 for $Q(.05)$. All higher horizontal lines have similar interpretations.

than a non-parametric estimator. Figure 7.7 shows the standard error of $\tilde{Q}$ (solid lines) with the asymptotic standard deviation of the corresponding sample quantiles for sample sizes of 80, 20 and 10 (dashed lines). These are the sample sizes of the three phases in the simulated data.

The standard error for $\tilde{Q}$ is relatively constant for the length of the profile. It is slightly lower at the beginning of the profile, where the fit of the imputation model is based on more observations. The shape of the reference curves reflects the changing shape of the density across the profile. Note that quantiles are "harder" to estimate in the right tail of the distribution since the height of the density is lower. For quantiles greater than $Q(.25)$, the height of the density increases monotonically across the profile. Thus, the reference curves for these quantiles are decreasing monotonically across the profile. For $Q(.10)$, the reference curves decrease for the first few inches and then increase for the rest of the profile.

In the right tail of the distribution, $\tilde{Q}$ has a lower standard error than Q_{80} for most of the profile. However, as we move toward the left tail of the distribution, the standard error of $\tilde{Q}$ becomes larger than that of Q_{10} for most of the profile. In general, it appears that for quantiles that are "hard" to estimate, the semi-parametric estimate $\tilde{Q}$ performs much better than the non-parametric estimates. The model brings stability to the estimate. For quantiles that are relatively "easy" to estimate, the variability in $\tilde{Q}$ due to estimating the model causes it to perform worse than the sample quantiles.

7.6 Variance estimation

As discussed in Section 4.5.2, jackknife methodology may be appropriate for estimating the variance of $\tilde{Q}$. In Chapter 5, a delete-d jackknife variance estimator is used. While theoretical results for this estimator are beyond the scope of this dissertation, some simulation results are of note.

Clusters of size four are created by randomly grouping three sites in $\mathcal{S}$ with one site in either $\mathcal{F}$ or $\mathcal{L}$. Strata of two clusters are then formed randomly. For the simulated data, this results in ten strata with two clusters each. For each jackknife replicate, cluster k

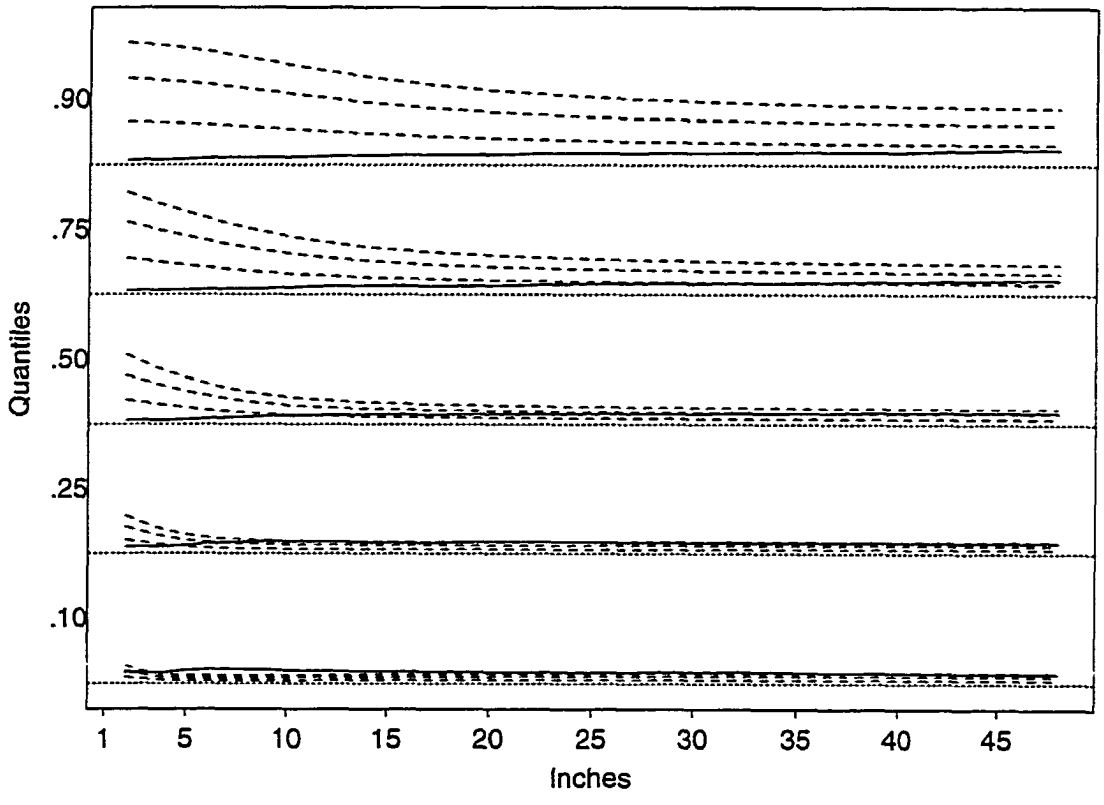

Figure 7.7 Standard error of $\tilde{Q}$ with asymptotic standard deviation of sample quantiles. Solid lines are standard errors of $\tilde{Q}$. Dashed lines are the asymptotic standard deviation of Q_{10}, Q_{20} and Q_{80}. Dotted lines separate segments of the plot. The scale is the same for all segments.

from stratum j is deleted. The remaining cluster in stratum j is duplicated to replace the deleted cluster. A quantile estimate, $\tilde{Q}^{(jk)}$, is calculated from the data with the jkth cluster removed. The jackknife variance estimator is then given by

$$v_{JK}(\tilde{Q}(p)) = \sum_{j,k} 0.5 \left(\tilde{Q}^{(jk)} - \tilde{Q}^{(j\cdot)}\right)^2, \tag{7.1}$$

where

$$\tilde{Q}^{(j\cdot)} = 0.5 \left(\tilde{Q}^{(j1)} + \tilde{Q}^{(j2)}\right). \tag{7.2}$$

For the original estimator and model, v_{JK} is calculated for nine quantiles. Figure 7.8 shows estimated standard errors from jackknifing and the empirical standard deviation of $\tilde{Q}$ for nine quantiles. The jackknife standard error is the average of 10 replications, while the empirical standard deviation is calculated from 100 replications. Only 10 replications were used for jackknifing because of computing time. Aside from the computing time, v_{JK} appears to perform very well.

In an effort to reduce computing time, a simplified jackknife variance estimator was investigated. For the full data set, the calibration and imputation models are fit. These estimates will not be recomputed for jackknife replicates. For each jackknife replicate, we remove all items in the imputed data set that were generated by elements in the jkth cluster. A variance estimator, $v_{JK,2}$ is calculated exactly as in (7.1) replacing $\tilde{Q}^{(jk)}$ with $\tilde{Q}^{(jk,2)}$. That is

$$v_{JK,2}(\tilde{Q}(p)) = \sum_{j,k} 0.5 \left(\tilde{Q}^{(jk,2)} - \tilde{Q}^{(j\cdot,2)}\right)^2. \tag{7.3}$$

This estimator essentially ignores the variability in estimating the coefficients of both the calibration and imputation models. Figure 7.9 shows $\sqrt{v_{JK,2}}$ versus the empirical standard deviation of $\tilde{Q}$. Again, the jackknife estimate is averaged over 10 replications. The estimator $v_{JK,2}$ appears to severely underestimate the variance in the left tail of the distribution. Although, computing time is greatly reduced, this estimator may be badly biased.

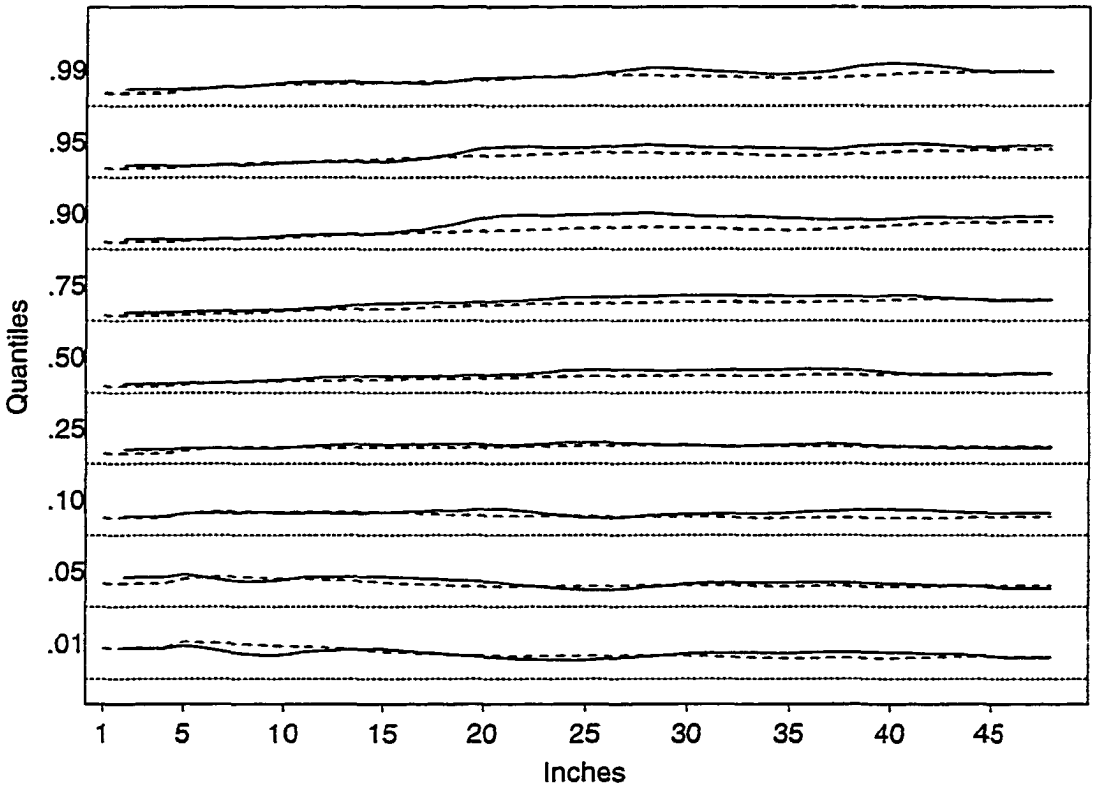

Figure 7.8 Jackknife standard error estimates (solid line) and empirical standard deviation of $\tilde{Q}$. Horizontal dotted lines represent a standard error of zero for each quantile and serve to separate portions of the graph. The scale is the same for each portion of the graph.

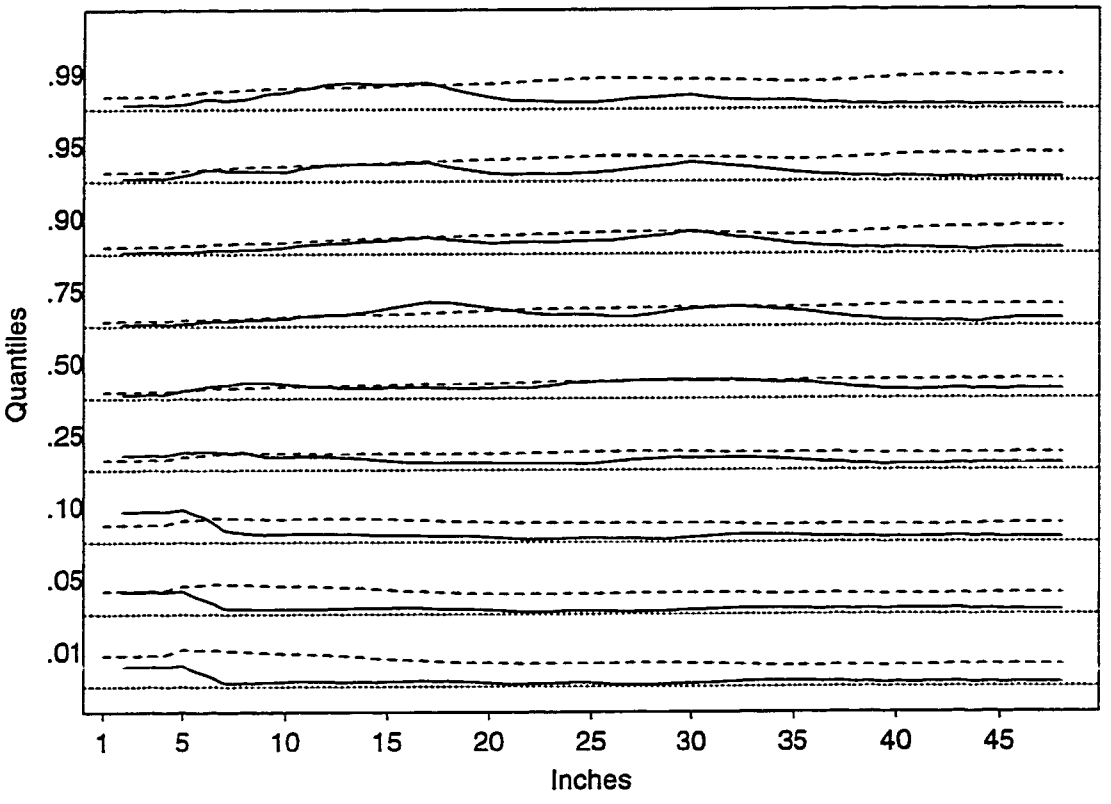

Figure 7.9 Simplified jackknife standard error estimates (solid line) and empirical standard deviation of $\tilde{Q}$. Horizontal dotted lines represent a standard error of zero for each quantile and serve to separate portions of the graph. The scale is the same for each portion of the graph.

7.7 Conclusion

The imputation approach requires fairly mild modeling assumptions and offers computational simplicity relative to the Bayesian approach. However, if we can verify that the assumptions of the hierarchical model hold even approximately, we have a useful explicit model. Posterior distributions of all parameters can be obtained. These parameters are also related to other variables of interest in the project. Thus, the Bayesian approach can provide a comprehensive framework for estimation in this project.

The hierarchical model can be used to simulate data which is analyzed using the imputation approach. The calibration step in the imputation approach appears to be a large source of bias. Simulations suggest that random calibration would improve the bias of $\tilde{Q}$. The variance of $\tilde{Q}$ is reasonable when compared to that of sample quantiles. While there is some loss of efficiency for quantiles where the density is high, the gains are considerable for quantiles with low values of the density.

The simulation also indicates that jackknife methodology may be appropriate for estimating the variance of $\tilde{Q}$. A version of a delete-d jackknife appears to perform quite well. However, the jackknife methodology needs more investigation.

8 CONCLUSION

The work for this dissertation was motivated by the MLRA 107 pilot project for updating soil surveys in western Iowa. A multi-phase sampling design was implemented in the pilot project. A major contribution of this sampling design is that data collected under this design can be used for a broader suite of statistical analyses than in a more traditional soil survey update. Statistically defensible estimates of distributional quantities such as means, percentiles, or parametric distributions can be generated along with estimated standard errors. The database resulting from this approach also supports geographically-linked modeling efforts.

We are interested in obtaining estimated quantile profiles for laboratory determinations of soil texture data collected under the multi-phase design. The sample was designed to take advantage of auxiliary information in the form of field measurements of texture. Chambers and Dunstan (1986) introduced a distribution function estimator (CDE) which attempts to do this. An asymptotic variance expression for a quantile estimator based on the CDE is derived (Chapter 4). Possible extensions of the CDE to three-phase samples are also suggested.

8.1 Analysis approaches

Two approaches to estimating soil texture quantile profiles are investigated. Texture is a three-dimensional vector whose components must sum to one. We transform the texture vectors to the two-dimensional real number space in order to simplify modeling. In both estimation approaches, the log-ratio transformation advocated by Aitchison (1986) is used.

In the first approach (Chapter 5), predicted values are calculated for missing laboratory profiles in two steps referred to as calibration and imputation, respectively. In the calibration step, a prediction from a linear regression model is used to impute missing transformed laboratory values from transformed field values. In the imputation step, we modify the procedure of Chambers and Dunstan (1986). In this step, several imputed profiles of transformed laboratory values are calculated for each site. Each imputed value is a prediction from a linear model plus a fitted residual from the model. Data are back-transformed to the original scale and a weighted empirical distribution function is used to calculate quantile estimates at each inch.

Assumptions of the calibration and imputation models appear to be reasonably satisfied using typical diagnostic methods for regression. However, comparing estimated quantile profiles to observed data indicates that this methodology may produce estimated distributions which are too peaked. This phenomenon is also seen in a small simulation study (Chapter 7). A random calibration step is proposed to combat the bias. This random calibration step is related to the generalized three-phase CDE proposed in Section 4.6.

A delete-d jackknife variance estimator for estimated quantile profiles under this approach is proposed. For each jackknife replicate, a cluster consisting of three phase 1 sites and one phase 2 or phase 3 site is removed. While theoretical results for the consistency of this estimator are not given, it appears to work well in simulations. The implementation of the delete-d jackknife seems intuitively reasonable, although more investigation is needed.

In the hierarchical modeling approach (Chapter 6), the field measurements are modeled as a function of laboratory measurements. Transformed laboratory measurements for each horizon are assumed to follow a normal distribution with a mean and variance which depend on the master horizon designation. A Markov chain is used to model the transitions between horizons. The quantile profiles are defined as a function of parameters in the model.

In a special case of the model, maximum likelihood estimates are available. However,

in general, a MCMC technique can be used to simulate a sample from the full posterior distribution of all parameters. The simulated draws can be used to approximate the marginal posterior distribution of any parameters of the model or of any function of the parameters, e.g., quantile profiles. The draws can also be used to estimate parameters of the posterior distribution, such as means, variances and posterior intervals. Thus, inference derived from the Bayesian approach can be more flexible than that available with the estimated variances of the imputation approach.

Posterior predictive assessment is used to evaluate the fit of the model. In particular, we find that the horizon profile model may be inappropriate. We suggest an improvement of the model in which the transition probabilities of the Markov chain are allowed to change with depth. A simple extension with this property was implemented and seems to improve the fit of the model. However, a more general extension is proposed in which the transition probabilities are modeled as a smooth function of depth. This extension has not yet been implemented.

The hierarchical model may also be useful for other analysis objectives of the pilot project. Many of the variables are horizon-based measurements which can be modeled similarly to the field and laboratory texture measurements. Other variables of interest can be defined directly as functions of the parameters of the horizon profile model. A comprehensive model for many variables would allow a unified analysis approach for the pilot project.

8.2 Future work

Some extensions for the hierarchical model were suggested in Section 6.9. These extensions may provide a better fit to the data and may be useful for other variables. In particular, other transformations of the texture vector may be used and a heterogeneous Markov chain model for the horizon profiles may be appropriate. Thus, we plan to further develop and apply the hierarchical model to data collected for the soils project.

Multi-phase designs are common in environmental applications of survey sampling. We plan to investigate results for extending the CD estimator to a three-phase data

structure. For both the standard CD quantile estimator and extensions, we are interested in studying the implementation of a delete-d jackknife for multi-phase data and the consistency of such a variance estimator.

BIBLIOGRAPHY

Abbitt, P. J., Breidt, F. J., and Nusser, S. M. (1997). A nonlinear two-phase predictor for soil survey updates. In *ASA Proceedings of the Section on Survey Research Methodology*, pages 657–650.

Abbitt, P. J., Goyeneche, J., and Schumi, J. A. (1998). An approach to estimating clay profile distributions. In *ASA Proceedings of the Section Survey Research Methodology*, pages 372–377.

Abbitt, P. J. and Nusser, S. M. (1995). Sampling approaches for soil survey updates. In *ASA Proceedings of the Section on Statistics and the Environment*, pages 87– 92.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52:200–203.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37:577–580.

Billingsley, P. (1995). *Probability and Measure*. Wiley.

Chambers, R. L., Dorfman, A. H., and Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, 79:577–582.

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88:268–277.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73:597–604.

David, H. A. (1981). *Order Statistics (Second Edition)*. Wiley.

Dorfman, A. H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *The Australian Journal of Statistics*, 35:29–41.

Dorfman, A. H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21:1452–1475.

Dunstan, R. and Chambers, R. L. (1989). Estimating distribution functions from survey data with limited benchmark information. *The Australian Journal of Statistics*, 31:1–11.

Francisco, C. A. and Fuller, W. A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19:454–469.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (disc: P760-807). *Statistica Sinica*, 6:733–760.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (disc: P483-501, 503-511). *Statistical Science*, 7:457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *The Annals of Mathematical Statistics*, 42:1957–1961.

Goyeneche, J. J. (1999). *Estimation of the Distribution Function using Auxiliary Information.* PhD dissertation, Iowa State University.

Iyengar, M. and Dey, D. K. (1998). Box-cox transformations in bayesian analysis of compositional data. *Environmetrics*, 9:657–671.

Kuk, A. Y. C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80:385–392.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In *ASA Proceedings of the Section on Survey Research Methods*, pages 280– 285.

Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, 73:194–196.

McCarthy, P. J. (1965). Stratified sampling and distribution-free confidence intervals for a median. *Journal of the American Statistical Association*, 60:772–783.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10:462–474.

Rao, J. N. K., Kovar, J. G., and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77:365–375.

Sarndal, C.-E., Swensson, B., and Wretman, J. (1991). *Model Assisted Survey Sampling.* Springer-Verlag.

Sedransk, J. and Meyer, J. (1978). Confidence intervals for the quantiles of a finite population: Simple random and stratified simple random sampling. *Journal of the Royal Statistical Society, Series B, Methodological*, 40:239–252.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag.

Smith, P. and Sedransk, J. (1983). Lower bounds for confidence coefficients for confidence intervals for finite population quantiles. *Communications in Statistics, Part A - Theory and Methods*, 12:1329–1344.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80:1026–1031.

Tierney, L. (1994). Markov chains for exploring posterior distributions (disc: P1728-1762). *The Annals of Statistics*, 22:1701–1728.

Wang, S. and Dorfman, A. H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83:639–652.

Wu, C. (1998). *The Effective Use of Complete Auxiliary Information From Survey Data.* PhD dissertation, Simon Fraser University.

ACKNOWLEDGEMENTS

This research was supported in part by Cooperative Agreement No. 68-3875-5-72 between the USDA NRCS Soils Division and the Iowa State University Statistical Laboratory.

I would like to thank both of my advisors and all of the members of my committee: Dr. Sarah Nusser, Dr. F. Jay Breidt, Dr. Hal Stern, Dr. Tom Fenton and Dr. S. N. Lahiri. Thanks to Dr. Sarah Nusser for guiding my professional development and for many valuable professional opportunities over the past five years. Thanks to Dr. F. Jay Breidt for guidance in many areas of my research and for patience with my programming skills. Thanks to Dr. Hal Stern for advice on development and assessment of the hierarchical model. Thanks to Dr. Wayne Fuller for much guidance in the art of analyzing real data, particularly the soils data. Thanks to Dr. Tom Fenton, Patrick Cowsert and Sam Steckly for teaching me about soils.

To Deanne, Schuckers, Johnny and my other contemporaries: thanks for listening to my failures and my successes and for sharing yours with me. Special thanks to my "study buddy" Deanne. From a leisurely summer of afternoon bike rides, to a not-so-leisurely summer of late night dissertating, we've shared much here in Ames. You know my door is always open for a mid-day office visit, even if it's only a virtual one!

Thanks to my family, for always having blind faith in me. Especially to my grandmother for teaching me that the shortest distance between two points is a straight line; to my brother for teaching me that the shortest distance doesn't always take you where you want to go; to my dad for teaching me to never settle for less than my best; and to my mother for her support in all things, for leading by example and for being the strongest person I know.

To my husband, Jamie, who provides never-failing moral support and much needed perspective: I know you have made significant sacrifices while waiting for me to finish my degree and great efforts to make life easier for me during the final months of my dissertation. At the very least, I owe you one cross-country motorcycle trip.